Molecular diagnostics of aleutian mink disease virus: applied use of next generation sequencing and phylogenetics

Hagberg, Emma Elisabeth

Publication date:
2017

Document Version
Publisher's PDF, also known as Version of record

MOLECULAR DIAGNOSTICS OF ALEUTIAN MINK DISEASE VIRUS

APPLIED USE OF NEXT GENERATION SEQUENCING AND PHYLOGENETICS

Ph.D. thesis by Emma Elisabeth Hagberg, DVM

Submitted: 31st of January 2017

Kopenhagen Fur | DTU Bioinformatics | Innovation Fund Denmark

Supervisor-team

Academic supervisor
Anders Gorm Pedersen, MSc., Ph.D., Professor
Center for Biological Sequence Analysis
Technical University of Denmark, Institute for Bioinformatics
Kemitorvet 208, DK-2800 Lyngby

Business supervisor
Anders Krarup, MSc. Ph.D., Head of development laboratory
Kopenhagen Diagnostics, Kopenhagen Fur
Langagervej 60, DK-2600 Glostrup

Academic co-supervisor
Lars Erik Larsen, DVM, Ph.D., Professor
National Veterinary Institute
Technical University of Denmark
Bülowsvej 27, DK-1870 Frederiksberg

Assessment Committee

Nicola Decaro
Professor, DVM, PhD, DipECVM
Vice director and research delegate at the Department of Veterinary Medicine
Università degli Studi di Bari, Italy

Thea Kølsen Fischer
Professor, MD, DMSc, MPG
Head of Virology Surveillance and Research, Director of National WHO Reference Laboratories for Influenza, Measles, Polio and Rubella
Statens Serum Institut, Denmark

Thomas Sichertiz-Pontén
Professor, MSc, PhD
Group leader Metagenomics, DTU Bioinformatics
Technical University of Denmark, Denmark

PREFACE AND ACKNOWLEDGEMENTS

The work presented in this thesis is the result of an industrial PhD-project performed in collaboration with Kopenhagen Fur, the Department of Bioinformatics (former Systems Biology) Technical University of Denmark (DTU), and the National Veterinary Institute (DTU Vet). The project was carried out between August 2013 and January 2017, and was funded by the Danish Fur Breeders' Association, the Innovation Fund Denmark (grant number: 1355-00106A), and Kopenhagen Fur.

Linking the project to the industry and how Kopenhagen Fur fits into the ecosystem

Denmark accounts for 30-40% of the production of mink pelts, and currently there are approximately 1400 mink farms in the country. Most of the farms are members of Kopenhagen Fur, a shareholder association founded in 1930 with the aim to protect the Danish mink industry's interests. With an annual turnover of DKK 10.9 billion (2015/2015) and an estimated market share of 60%, Kopenhagen Fur is the world's largest fur auction house (www.kopenhagenfur.com). In addition to grading, sorting, and selling raw skins at five yearly auctions, Kopenhagen Fur provide consulting services to its members in areas related to farm operations, e.g. animal welfare, breeding, nutrition, disease control, economy, and a diagnostic service. In 2014 mink pelts was the third largest Danish export commodity with an estimated value of EUR 1.5 billion (www.agricultureandfood.dk).

Unexpected events

During the season of 2015-2016 (the second year of the project), Denmark was confronted with the largest plasmacytosis outbreak in history, and I had a six months leave from the PhD-project and worked full-time at Kopenhagen Fur. I facilitated and coordinated the collection of samples from all over the world for a comprehensive global AMDV-study. The samples collected in relation to the outbreak were used for the establishment of a global AMDV map based on partial NS1 gene sequencing, and I have co-authored two papers and a conference abstract based on these data (see overview of papers). During this period I was also responsible for training a laboratory technician in working with DNA-, PCR-, and real-time PCR techniques.

Steep learning curve

When I began working on this project in August 2013, I was aware that computers had a so-called command-line, or at least I had heard about it. In other words: I was a total beginner at bioinformatics and I was about to initiate the largest genomic study performed within AMDV research using some of the latest technologies and analyses available. On the bright side, I had technical flair and the drive to throw my self out there and learn. Three years later, I appreciate the excitement of being able to open and close documents in the computer without using the touchpad.

Acknowledgements

Several people have been involved throughout this project, and I would like to take the opportunity to thank them personally. First up is my main supervisor Prof. Anders Gorm Pedersen, who I would like to thank for his enthusiasm and willingness to share knowledge. It has been a privilege to learn from someone so good at explaining complex topics and with a true passion for everything (?) Bayesian.

I would like to thank my company supervisor Anders Krarup for his patience, pragmatic approach to science, for taking time to answer all of my questions, and for discussing the existence of human

nature and how the world could become a better place. A word of gratefulness goes to my colleagues at Kopenhagen Fur, for providing valuable background information about the industry, the Danish farmers, and the virus. Leif Bruun, former director and CSO of Kopenhagen Diagnostics, Kopenhagen Fur, is acknowledged for his visionary thinking and for supporting the initiation of this project.

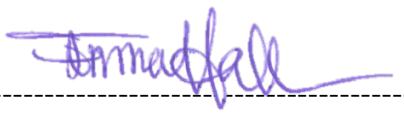
My co-supervisor, Prof. Lars Erik Larsen, is thanked for his encouraging and friendly attitude, for sharing his expertise about virology and AMDV, and for inviting me to his Virology group at NVI Frederiksberg. In total, I spent around six months in his group, working both with my own data, but also with real-time PCR development with the group. Which leads me on to senior researcher Charlotte Hjulsager and research assistant Pia Ryt-Hansen (now PhD-student at NVI). Both are gratefully thanked for their friendly mind-sets and excellent collaboration during the AMDV outbreak in 2015/2016.

In May 2015 I got the opportunity to visit Tanja Stadler's research group Computational Evolution (cEVO) at the Department of Biosystems Science and Engineering ETH Zürich in Basel (Switzerland). I would like to thank the group members in general for their constructive input, and a special thank you to Louis du Plessis (now in Oliver Pybus group in Oxford), Denise Kühnert, David Rasmussen, Nicola Müller, and Tanja Stadler.

Many thanks to the Danish e-infrastructure Centre (DeiC) for granting my project as a pilot-project, where I received computing hours and media exposure. Many thanks to project-manager Ali Syed and his staff at the National Life Science supercomputer 'Computerome'. Having access to their facilities significantly improved the computing time in the project, and Alis positive and solution-oriented approach was invaluable. I would like to thank the former NVI PhD-student Ulrik Fahnøe (now at Hvidovre Hospital) for helping with the initial development of the data analysis pipeline and for setting up my MacBook Pro to run the required bioinformatics software. And also, many thanks to two special veterinarians: Haiko Koenen (DAC Zuidoost), for helping me collect samples in The Netherlands, and for engaging in discussions about politics, diagnostics, and AMDV; and Ylva Naeve (Ugglarps Gård Hästkliniken), for proof-reading my thesis.

Finally, I would like to thank my family and my friends for staying around all of these years, despite long and busy hours, and times with a lot of travelling. However, the without doubt most important source of advice and support during this process has been my love, Christian Raaby. His support and encouragement has helped me to remain focused and able to see the light at the end of the tunnel (i.e. the 31st of January).

Best wishes,



Emma Hagberg
Copenhagen, 31st of January 2017

SUMMARY

Aleutian Mink Disease virus (AMDV) is a parvovirus causing Aleutian Mink Disease (AMD), often referred to as plasmacytosis. It is a systemic infection affecting mink of all ages, and is globally the most important pathogen impacting mink farming. In Denmark AMDV has since 1999 been monitored by a national control program, which is based on serological screening of all animals and encourages infected farms to stamp out. Historically there has been no consensus about which genomic region of the virus to analyse e.g. in relation to surveillance, and most previous studies in this regard, have been based either on partial or entire genes, or on pure epidemiological data. Thus, when initiating this project, little was known about AMDV's total genomic diversity and how the virus was spread between farms.

Recent advances in the field of molecular diagnostics have made high-throughput tools such as next generation sequencing cheaper and more easily available. Whole genome sequencing and advanced phylogenetic analyses have successfully been applied to describe the molecular evolution and transmission patterns for viruses such as Foot and Mouth Disease Virus (FMDV), Ebola, and avian influenza virus, however not previously for AMDV. The overall aim with this thesis was to investigate if next generation sequencing and phylogenetic analyses of full-length isolates could improve our understanding of the total genomic diversity and evolution of AMDV. Additionally, we wanted to evaluate if this knowledge could contribute to the elucidation of AMDV transmission between farms and improve molecular diagnostics.

During the first phase of this project a method for performing whole genome sequencing of AMDV was developed. This protocol enabled the sequencing of a large number of *in vivo* infectious AMDV isolates and provided the necessary dataset to act as foundation for the remaining analyses in the thesis. The first original paper (Manuscript 1) describes this protocol.

Manuscript 2 is a proof-of-concept study which demonstrated the advantage of using the whole genome sequence approach, compared to the in Denmark traditionally used partial NS1 gene sequencing, for the elucidation of transmission pathways between farms. The study was performed on samples from a small local AMDV outbreak, and clearly illustrated that the phylogenies based on partial NS1 gene sequencing were uninformative and could not be used for determining transmission pathways, even in the light of supporting epidemiological data. The whole-genome approach on the other hand, confirmed the epidemiological hypothesis about the direction of spread.

In Manuscript 3, the methodologies from Manuscript 1 and 2 were applied to generate the to-date most comprehensive phylogenetic and genetic analysis of full-length AMDV isolates, composed of more than 200 field strains. The study shed light on the diversity and evolutionary behaviour of two distinct AMDV strains, in addition to providing the first robust evolutionary rate-estimates. Altogether, the work presented in this thesis provides a contribution to the molecular diagnostics of AMDV, enables us better to understand the virus' evolutionary behaviour in the context of mink farming, and is anticipated to be of value for more accurately tracing back in time the emergence of future outbreaks.

RESUMÉ

Aleutian Mink Disease virus (AMDV) er et parvovirus, der i mink af alle aldre forårsager en systemisk infektion kaldet Aleutian Mink Disease. Sygdommen er også kendt som plasmacytose, og er på verdensplan den mest betydningsfulde indenfor minkavl. I Danmark har man siden 1999 overvåget AMDV ved et nationalt kontrolprogram, der baseret på serologisk screening identificerer og opfordrer inficerede farme til at slå bestanden ned. Historisk set, har der ikke været nogen konsensus vedrørende hvilken del af virussens genom der er blevet brugt til smitte efterforskninger, hovedparten har enten fokuseret på hele eller dele af gener, eller på rent epidemiologiske data. Derfor var der, ved opstarten af dette projekt, en relativt begrænset viden om AMDV's totale genetiske diversitet og dens spredning mellem farme.

Indenfor de seneste årtier, har der været mange teknologiske fremskridt indenfor molekylærbiologien, og high-througput metoder som næste generations sekventering (NGS) er blevet billigere og mere tilgængelige. Fordelene ved, at sekventere og analysere hele virale genomer i evolutions studier er, at dette giver mere genetisk information, og brugbarheden af dette i smittesporings øjemed, er blevet vist for patogener som f.eks. mund og klovsyge, Ebola, og aviær influenza. Det overordnede formål med denne afhandling var, at undersøge om NGS og fylogenetiske analyser af hele AMDV-genomet, kunne bidrage til et øget kendskab til virussens totale molekylære diversitet og evolution, samt at undersøge om denne viden kunne bruges til at udlede spredningsmønstre mellem farme og forbedre den molekylære diagnostik.

Under projektets første fase, blev en metode til at fuldgenomsekventere AMDV udviklet. Protokollen muliggjorde sekventeringen af et større antal *in vivo* infektiøse AMDV stamme, og dermed danne datagrundlaget for de resterende dele af projektet. Metoden er beskreven i Manuskript 1.

Manuskript 2 er et proof-of-concept studie, hvor fuldgenom sekventering sammenlignedes med partiel sekventering af NS1 genet som benyttes som standard i Danmark. Dette studie viste fordelen ved at bruge fuldgenomer i forhold til det partielle NS1 gen, i forbindelse med smittesporing. Studiet var baseret på et mindre dataset bestående af prøver med oprindelse i et lokalt AMDV udbrud, og viste tydeligt at fylogenierne baseret på partiel NS1 ikke kunne bruges til at udlede smitteveje. De fuldgenom-baserede analyser konfirmerede derimod den epidemiologiske hypotese vedrørende smittens retning.

I Manuskript 3 blev metodologierne fra Manuskript 1 og 2, benyttet til at generere mere end 200 fuld-genom sekvenser til den til dags dato mest omfattende studie af fuld-længde AMDV isolat. Studiet belyste virussens evolution og diversitet, og er den første rapportering af robuste estimater for den molekylære clock-rate. Resultaterne præsenteret i denne afhandling er et bidrag til den molekylære diagnostik af AMDV, har skabt mere viden om virussens evolution, og forventes at skabe værdi i forbindelse med datering og opklaring af smitteveje ved fremtidlige sygdoms udbrud.

LIST OF PAPERS

The following papers were produced as the result of this thesis and constitute a part of it:

Manuscript 1: A fast and robust method for whole genome sequencing of the Aleutian Mink Disease Virus (AMDV) genome.
Authors: Hagberg, E.E., Fahnøe, U., Larsen, L.E., Dam-Tuxen, R., Krarup, A., Pedersen A.G.
Status: published in Journal of Virological Methods (JVM), April 2016.

Manuscript 2: Evolutionary analysis of whole genome sequences from Aleutian Mink Disease Viruses confirms inter-farm transmission.
Authors: Hagberg, E.E., Pedersen A.G., Larsen, L.E., Krarup, A.
Status: accepted for publication in Journal of General Virology (JGV), March 2017.

Manuscript 3: Genetic analysis of the entire genome of Aleutian Mink Disease Virus determines its evolutionary rate and confirms bottleneck due to control program.
Authors: Hagberg, E.E., Larsen, L.E., Krarup, A., Pedersen, A.G.
Status: draft in preparation.

Manuscript 4: Development of a real-time PCR assay for detection of Aleutian Mink Disease Virus.
Authors: Hagberg, E.E., Krarup, A., Hjulsager, C.K., Folsing, C.B., Pedersen A.G., Larsen, L.E.
Status: the project is still in process, and the preliminary results are written in the form of a manuscript draft.

The following papers have been co-authored by the PhD-student, but do not constitute a part of this thesis:

Manuscript 5: Outbreak investigation of Aleutian Mink Disease Virus (AMDV) using partial NS1 gene sequencing.
Authors: Ryt-Hansen, P., Hjulsager, C.K., Chriél, M., Struve, T., Pedersen, A.G., Larsen, L.E.
Status: submitted to Veterinary Microbiology, December 2016.

Manuscript 6: A global molecular characterisation of Aleutian Mink Disease Virus (AMDV).
Authors: Ryt-Hansen, P., Hagberg, E.E., Chriél, M., Struve, T., Pedersen, A.G., Larsen, L.E., Hjulsager, C.K.
Status: draft in preparation.

ABBREVIATIONS

AIC – Aikake information criterion
AMD – Aleutian mink disease
AMDV – Aleutian mink disease virus

BD – Birth-Death
BEAST – Bayesian evolutionary analysis using trees
BIC – Bayesian information criterion

CIEP – Counter Current Immune Electrophoresis

C_q – quantification threshold

E_n – effective population size

HRP – horseradish peroxidase

NJ – neighbour joining

KF – Kopenhagen Fur

MCMC – Markov Chain Monte Carlo
ML – maximum likelihood
MRCA – most recent common ancestor
MSA – multiple sequence alignment

N_e – effective population size
NJ – neighbour joining
NGS – next generation sequencing
NS – non-structural

ORF – open reading frame

PCR – polymerase chain reaction

QC – quality control

R_e – effective reproductive number
R_o – reproductive number

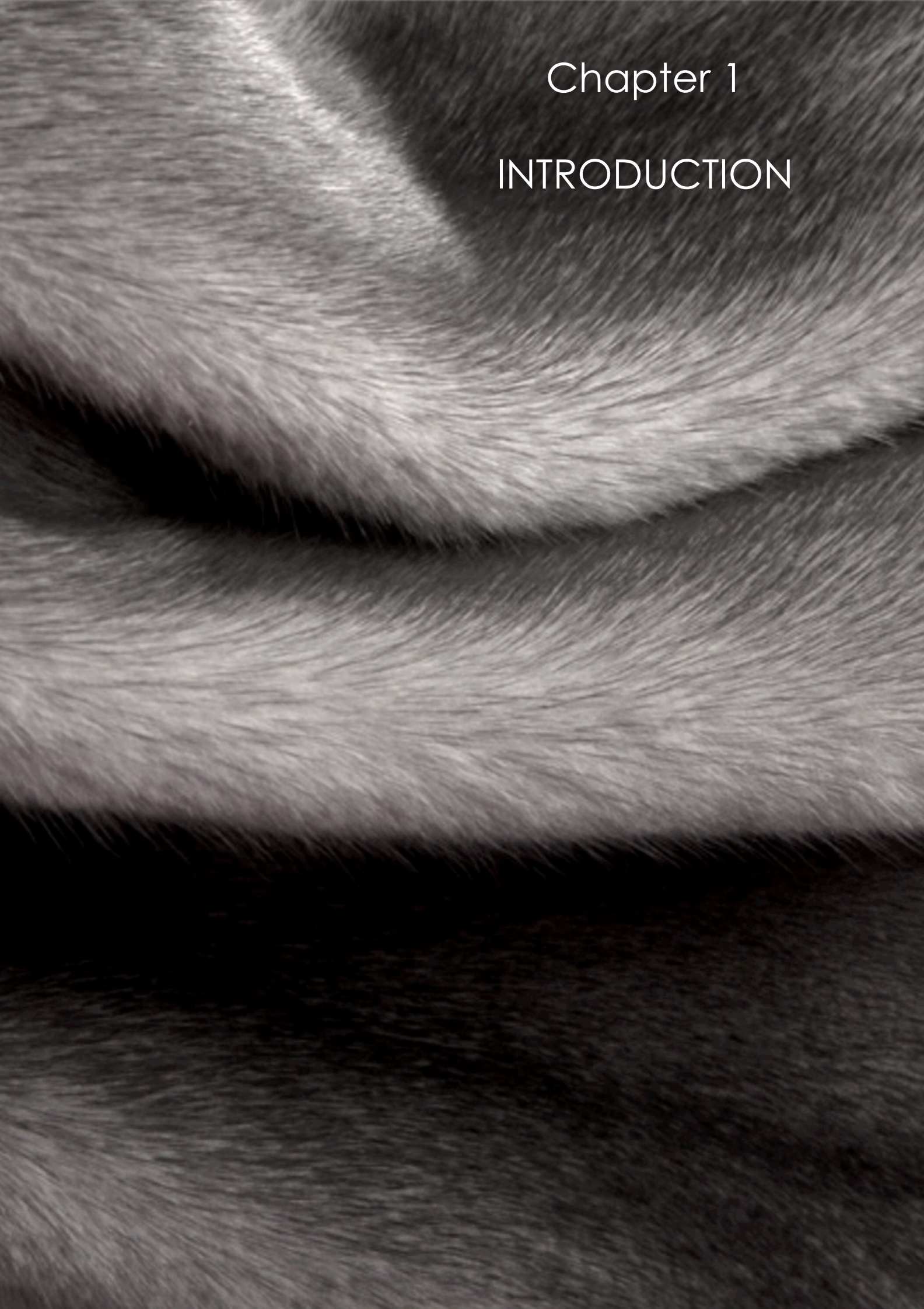
SIR – susceptible infected recovered
SNP – single nucleotide polymorphism
SNV – single nucleotide variant

UPGMA – unweighted pair group method using arithmetic mean

VP – viral (capsid) protein

WGS – whole genome sequencing/sequences

TABLE OF CONTENTS



Chapter 1

INTRODUCTION

1. INTRODUCTION AND OBJECTIVES

1.1. MINK FARMING AND THE FUR INDUSTRY

The approximately 1400 Danish mink farms account for 30-40% of the world production, and mink pelts are the third largest export commodities from Denmark to China (www.kopenhagenfur.com). On a global scale, the remaining pelts are mainly produced in Canada, USA, China, Holland, Poland, Finland, Russia, and to a smaller extent in countries such as Sweden, Italy, Norway, Greece, and Spain (fig. 1).

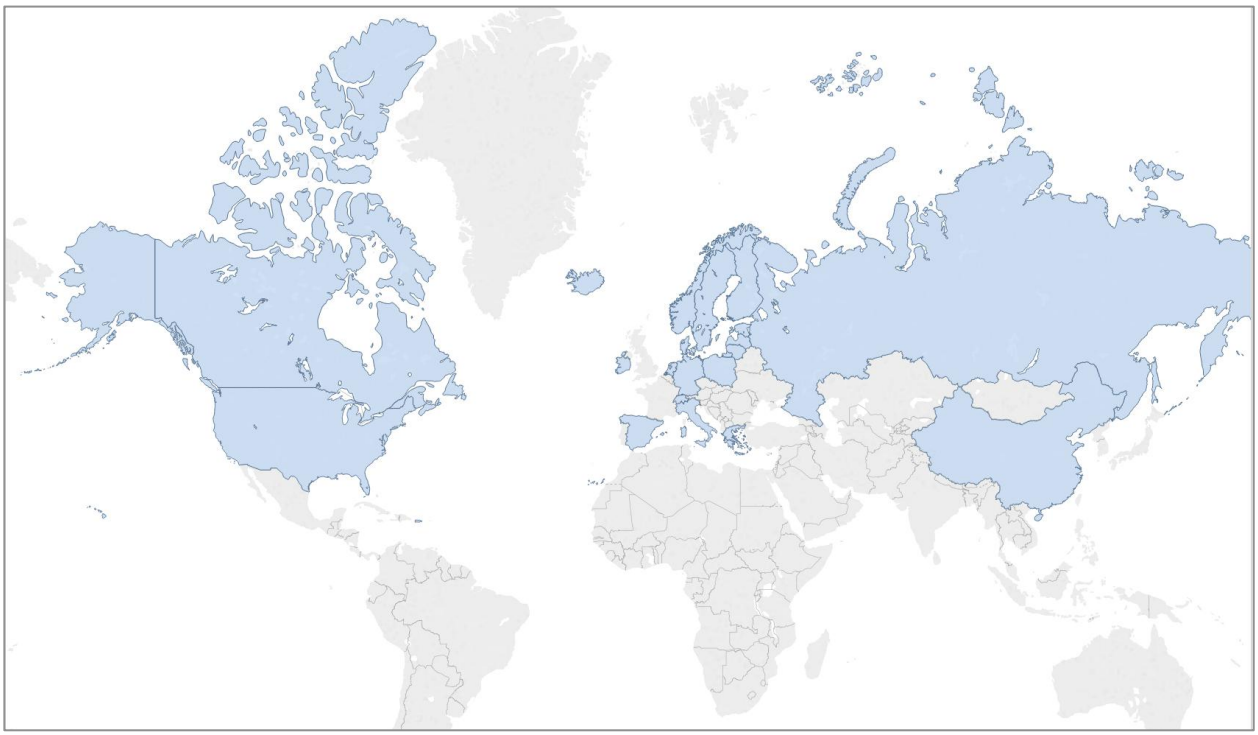


Figure 1. World map of mink-producing countries. Mink producing countries in light blue. Map created using Tableau v.9.2.

1.2. MINK FARMING AND PLASMCYTOSIS

Aleutian Mink Disease (AMD), often referred to as plasmacytosis, is caused by Aleutian Mink Disease virus (AMDV), and is globally considered the most important disease in mink. The various fur producing countries deal with AMD in different ways. For example, in Holland the disease is maintained "under control" by removal of infected animals identified by serological screening (chapter 2.3.2.), and similar strategies are employed in Canada and presumably also in China (personal communication, KF). Denmark has chosen an eradication strategy where the disease is monitored by

a national control program implemented in 1976 (Chriél 2000) and regulated by law since 1999 (Anon 2009). Briefly, the program requires member farms to conduct serological screening of their mink at regular intervals depending on the disease status of the region. Positive farms undergo more intensive monitoring and are encouraged to depopulate, followed by thorough cleaning and disinfection of the farm according to specific guidelines before repopulating. Due to the success of the control programme, AMD has been more or less eradicated in the country, except for in an endemic area in Northern Jutland. However, there have been relapses, such as the presumed foodborne outbreaks in Sole in 2002 (Willadsen 2003), where a significant number of farms were affected with AMDV. Intensive serological testing in addition to consistently stamping out infected farms rapidly contained the outbreak and few years later the situation once again was stable, and remained so until the season of 2015-2016, where the Danish mink industry experienced its largest AMD outbreak. In the summer of 2015 farms around the city of Holstebro, far south of the well-known endemic area (fig. 2), began to serologically test positive for AMDV (Ryt-Hansen, Hjulsager, et al. 2017). Efforts were made to limit the spread, but in December 2015 the first case on Zealand was confirmed. Once again control measures such systematic testing, stamping out, and limiting the movement of animals, equipment and people between farms contributed to gain control.

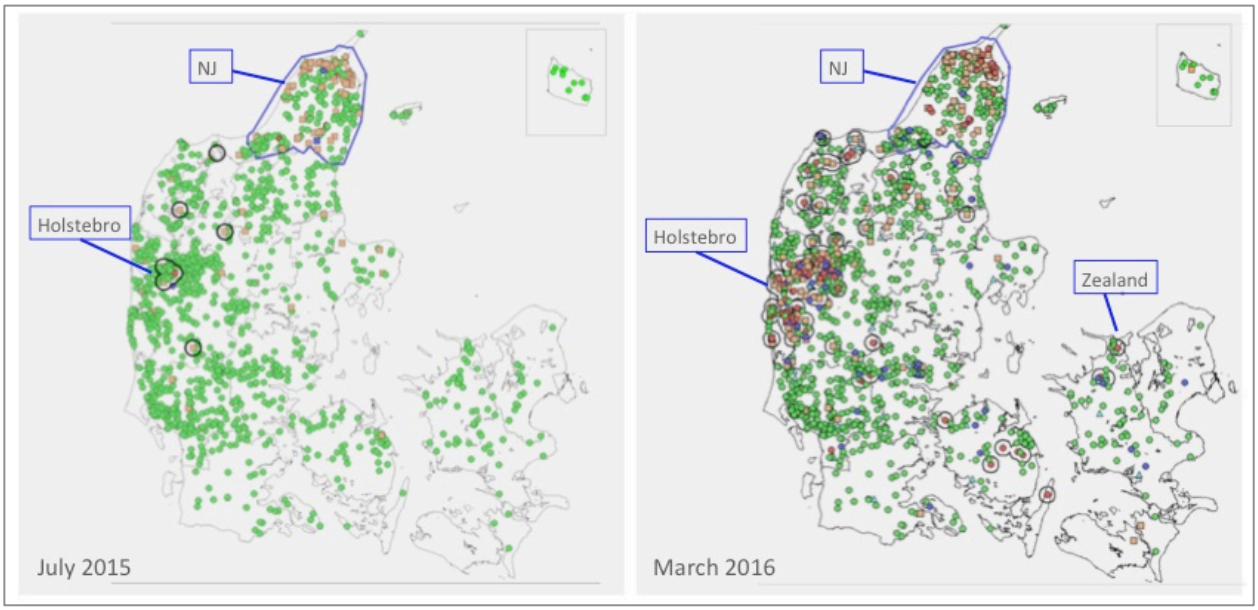


Figure 2. AMDV status in Denmark in July 2015 and March 2016. The 2015 map shows the endemic area in Northern Jutland (NJ) and the first cases in Holstebro. The 2016 map shows the status on March 2016. AMDV-free farms (green), AMDV-infected farms (red), AMDV-suspected farms (orange), depopulated farms (blue), and disease zones (blue lines). Underlying maps provided by Anne Berthelsen, Kopenhagen Fur.

There are several challenges related to stamping out animals on farm level, e.g. the infected carcasses need to be taken care of according to the regulations, and the farms need to be very thoroughly cleaned and disinfected to ensure there is no virus left. The latter can be very challenging as most farms have a myriad of cages and woodwork where the virus, which is very resistant to environmental conditions, can persist. There are furthermore emotional and economical aspects related to the stamping out process, such as motivating the farmers to euthanize the livestock they have dedicated years and sometimes generations to breed. The costs associated with cleaning the farm are often high, and identifying and buying new animals is difficult during larger outbreaks due the heavily increased demand for high quality breeding stock. Thus, managing AMDV is not only laborious but imposes large costs on the individual farmers and on the industry in general.

1.3. IDENTIFICATION OF RESEARCH OPPORTUNITIES

Literature surveys revealed gaps in our knowledge about AMDV. A prerequisite for developing efficient molecular tools is thorough knowledge of the pathogens entire genome, its variation within and between hosts, and its rates and mechanisms of evolution. This section aims to map out some of these gaps:

i) The benefits from adding more information

When this study was initiated in 2013, the relatedness between AMDV isolates had mainly been investigated based on pure epidemiological data (Themudo et al. 2011) or by analysing sequences originating either from partial (Jensen et al. 2011) or entire genes (Sang et al. 2012; Leimann et al. 2015; Knuuttila et al. 2015; Oie et al. 1996). The superiority of using whole genome sequencing for describing molecular evolution and for identifying transmission patterns had been illustrated in other viruses such as Foot and Mouth Disease Virus (FMDV) (Morelli et al. 2013), Ebola (Gire et al. 2014), and swine influenza virus (Watson et al. 2015), but not for AMDV.

ii) Collecting information from suitable proxies

Most knowledge about the entire AMDV genome and its replication was based on studies in AMDV-Goreham (AMDV-G), a cell culture adapted strain that is only infectious *in* vitro, and grown in a Crandall Rees Feline Kidney (CRFK) cell line. The *in vivo* pathogenic AMDV strains circulating on the farms, do on the other hand not replicate in cell culture (Bloom, Kaaden, et al. 1988). This has hampered in-depth experiments with field strains and also illustrates that AMDV-G might be a questionable model system.

iii) Capitalising on technological advances

The first pioneering studies that characterised the entire AMDV-G genome utilized traditional molecular methods, such as restriction fragmentation (Aasted 1980) and cloning (Aasted 1980; Alexandersen et al. 1988). More recent studies have generated whole genome sequences from AMDV field strains by amplifying the viral DNA in multiple overlapping PCR-fragments (Li et al. 2012; Xi et al. 2016) or by nested PCR (Canuti et al. 2016). All of the above-mentioned studies applied so-called "first generation" sequencing methods, and were altogether more labour-intensive and less suitable for efficient high throughput implementation e.g. in surveillance programmes. Next generation sequencing (NGS) is a powerful set of technologies that can increase data accuracy and generate additional layers of information. During the past decades NGS have become significantly cheaper, more readily available, and is suitable for high throughput applications. NGS has been successfully applied to characterise the entire genomes of other viruses, and the obtained knowledge has improved preventative measures of e.g. FMDV, Porcine Parvovirus (PPV), and Ebola (Escobar-Gutiérrez et al. 2012; Jakhesara et al. 2014; Kvisgaard et al. 2013; Gire et al. 2014). Phylogenetic analyses has even been used to aid in court decisions (Metzker et al. 2002).

1.4. OBJECTIVES AND AIMS

The working hypotheses for this project was that by learning more about the AMDV genome and its evolution, it would be possible to improve the molecular diagnostic tools, as well as to investigate if whole genome analyses could be used for elucidating transmission routes during disease outbreaks. To address these questions the project was structured into four phases:

I) The aim of the first phase was to develop a method to efficiently amplify and sequence the full genome of AMDV field strains using NGS. It was important that the method was easy to implement and validate in the laboratory. Manuscript 1 is the result of this work.

II) In the second phase the above-mentioned method was applied to viral strains originating from a small AMDV outbreak, and the resulting data created the base for a case study of the use of whole genome sequencing as a tool for outbreak investigation. Manuscript 2 describes this work.

III) In the third phase, the results and conclusions from the first and second phase were applied on a larger set of viral samples, and the total AMDV diversity and evolution was investigated. Manuscript 3 is the summary of this work.

IV) The knowledge from the previous phases could then, in the fourth phase, create the base for the development of improved molecular tools. Specifically, a real-time PCR assay for detection of AMDV was developed from the data generated in this project. Manuscript 4 summarises the status of this phase.

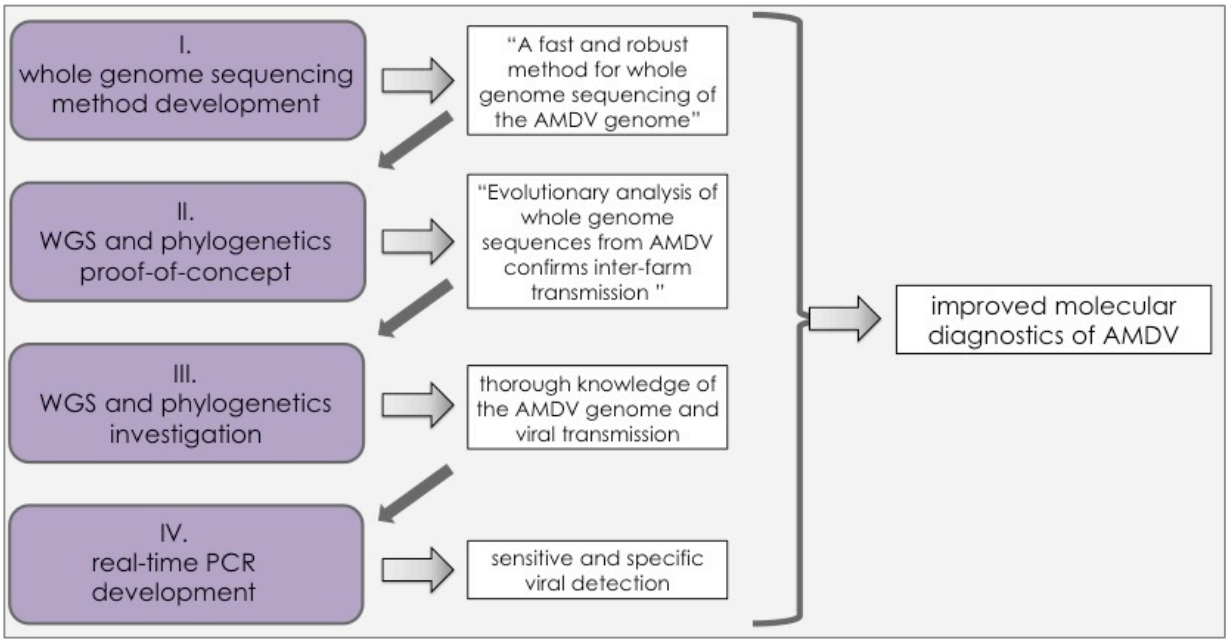
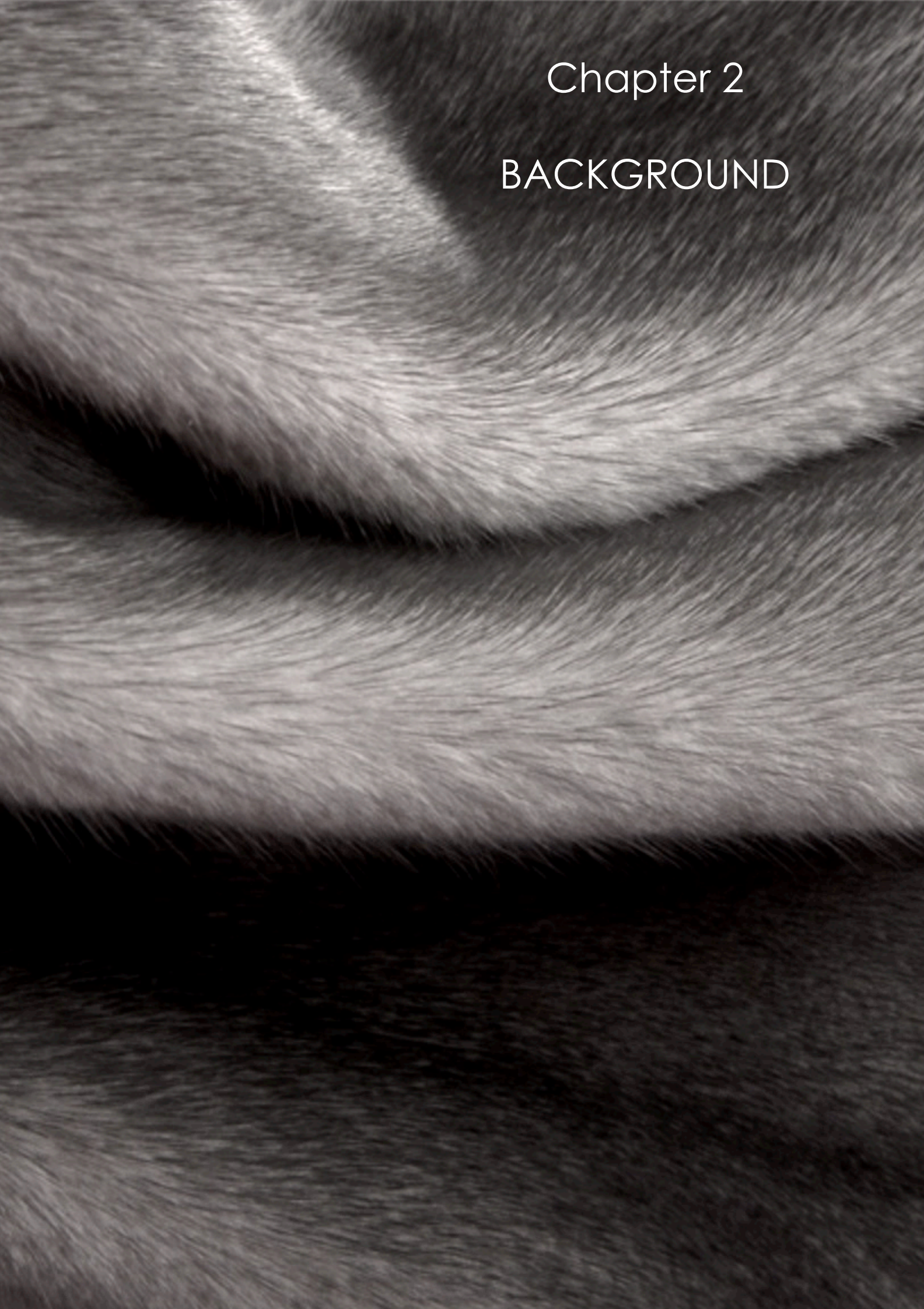


Figure 3. Project flow-chart. Illustration of the relationship between the project phases and how their output adds value during the next step.



Chapter 2

BACKGROUND

2. BACKGROUND

2.1. ALEUTIAN MINK DISEASE VIRUS

2.1.1. Biology and pathogenesis

Aleutian Mink Disease Virus (AMDV) was classified as belonging to the family *Parvoviridae* in the 1980's (Bloom et al. 1980), and was in 2014 further grouped into the subfamily *Parvovirinae* genus *Amdoparvovirus* and species *Carnivore amdoparvovirus 1* (Anon n.d.). Despite that most parvoviruses are host-specific they are classified according to their molecular properties and not their host-species of origin, creating a complex system of classification. Parvoviruses generally cause mild infections, however there are several parvoviruses of major veterinary importance, such as Porcine parvovirus (PPV), Canine parvovirus, Feline Panleukopenia virus, AMDV, and Mink enteritis virus (MEV) (Maclachlan et al. 2011), in addition to B19 - the most important parvovirus infecting humans (Fields et al. 2007).

Parvoviruses are characterised by their single-stranded DNA genome and a small (25nm diameter) non-enveloped viral capsid (fig. 4), making them durable and resistant to environmental factors (Maclachlan et al. 2011). Viral entry has been showed to occur through the respiratory, oral, or trans-placental routes (Broll & Alexandersen 1996). Few studies have specifically investigated the progression of an AMDV infection. However, in similarity with other viral infections an initial incubation period of 1-7 days has been demonstrated, followed by an increase in immunoglobulin G (IgG) and a transient viraemic phase (Jensen et al. 2015). AMD is characterised by a massive and harmful activation of the immune system, which results in hypergammaglobulinaemia and deposition of immune complexes in the vascular tissue.

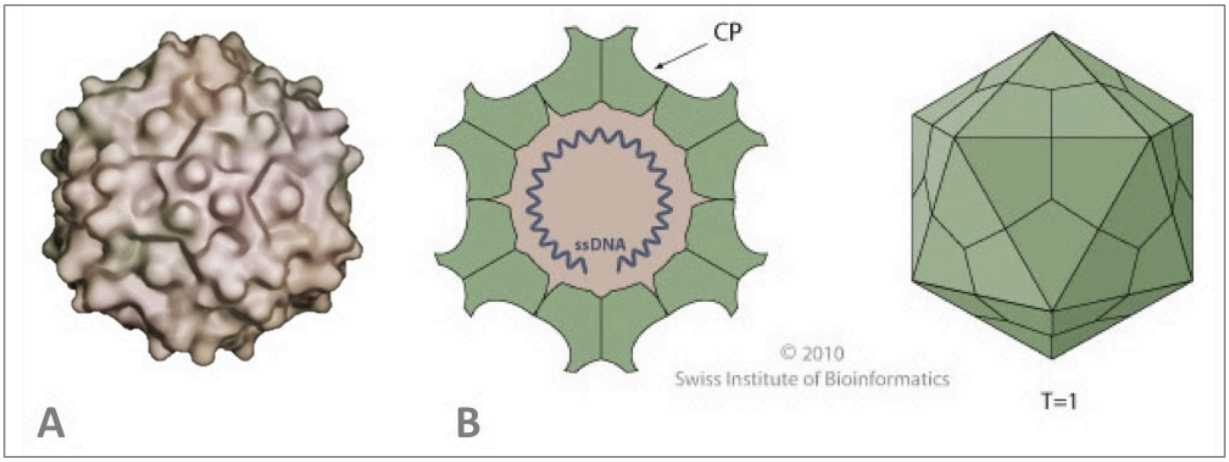


Figure 4. AMDV particle and capsid. Electro-micrograph of an AMDV particle (panel A) and a schematic illustration of a parvovirus capsid and its single stranded DNA strand (panel B on behalf of Swiss Institute of Bioinformatics).

AMDV is the causal agent of Aleutian Mink Disease (AMD), also referred to as plasmacytosis, a disease which mainly affect farmed mink (*Neovision vision)*, but has also been described in wild mink (Persson et al. 2015; Nituch et al. 2012), other *mustelidae* such as skunk (Nituch et al. 2015; Farid 2013) and otters, and in bobcats as well as raccoons (Farid 2013). The name plasmacytosis reflects a targeting of cells in the immune system of adult mink, and that the infection leads to an expansion of plasma cells resulting in excessive production of antibodies. Mink of all ages can be infected with AMDV, but due to a tropism for actively dividing cells the disease manifestation varies. The most common disease form in mink-kits is the so-called lung plasmacytosis, where the virus replicates in alveolar type 2 cells, and the disease is manifested as severe interstitial pneumonia with high mortality. The classic non-specific and chronic plasmacytosis is more common in adult mink, often with vague clinical signs (Jensen et al. 2015) such as reduced growth, decreased fertility, increased mortality, reduced fur quality with sprinkled white hairs (Farid & Ferns 2011), and a characteristic bulldog nose due to damaged blood vessels in the nasal mucosa (fig. 5). These clinical signs are mainly the results of systemic vascular conditions such as glomerulonephritis and arteritis. Currently there is no treatment or vaccine against AMD, and the infected mink either die due to organ failure or become persistently infected carriers transmitting the virus within and between herds (Decaro, Nicola et al. 2012).

Figure 5. Clinical signs of AMD. Sapphire coloured farmed mink: upper left, a healthy mink with shiny hair coat and awaken eyes, and upper right a mink with exudation around its eyes and a characteristic bulldog nose. Close-up of a healthy pelt (lower left) and a pelt with the plasmacytosis characteristic sprinkled white hairs (lower right).

2.1.2. The AMDV genome and its replication

The AMDV genome is typical for a parvovirus and consists of a single-stranded linear DNA (ssDNA) strand, is of approximately 4.800 nucleotides in length, and has tandem-repeated hairpin structures at the 5'- and 3'-ends (Bloom, Alexandersen, et al. 1988; Bloom et al. 1990). Characteristic for parvoviruses is that they do not encode their own DNA polymerase and therefore fully relies on utilizing the host cells replication machinery (Maclachlan et al. 2011). Hence, AMDV only replicates in dividing cells and exhibits tropism for e.g. hematopoietic precursors and lymphocytes in the spleen and lymph nodes. AMDV can additionally be found in organs from which there are attempts of clearance from the host, e.g. in the lungs, kidneys, and liver (Maclachlan et al. 2011).

Given the parvoviruses small genomes they have evolved means to maximise its usage, such as combining the same piece of DNA into several products (splicing), using multiple promoters,

alternative splice donor and acceptor sites, unusual initiation codons, and proteolytic cleavage (Berns 1990). The hairpin structures in the ends also contribute as initiation sites for the DNA polymerase (Maclachlan et al. 2011). AMDV's transcription profile is similar to that of MEV and other parvoviruses (Qiu et al. 2006), where during replication, six species of mRNA are produced through alternative splicing and alternative polyadenylation of a pre-mRNA generated by a single promoter. The result is three different splicing patterns, all of which accumulate during an infection. The coding regions contains two large open reading frames (ORF's); the left ORF coding for the non-structural (NS) proteins involved in gene regulation and replication, the right ORF coding for the viral capsid proteins (VP); and three smaller central ORF's (Alexandersen et al. 1988; Bloom, Alexandersen, et al. 1988; Hagberg et al. 2016).

2.1.2.1 The left ORF

The left open reading frame (ORF) is located between nucleotides 116-1975 in the AMDV-reference genome NC001662 (fig. 6). It encodes for the non-structural proteins: NS1, NS2, and NS3. NS1 is considered the major regulatory protein in parvoviruses and plays an important role in viral replication during infection (Fields et al. 2007; Gottschalck et al. 1994). Between amino acids 421-492 there is a GKRN-region (Gottschalck et al. 1994), which e.g. contains a purine binding site between amino acids 435-440, suggested to be important for the ATPase function, and for promoting conformational changes in mRNA affecting translation. The PKR-region between NS1 aa 623-625, previously proposed to be involved in nuclear localisation (Gottschalck et al. 1994). In the NS1 gene the highest degree of variability has been reported for N- and C-terminals, as compared between the cell culture adapted AMDV-G and the virulent AMDV-Utah (Gottschalck et al. 1994; Hagberg et al. 2016). The smaller and central ORF's code for the 3'-ends of the NS2 and NS3 genes, which are spliced together with the main bulk of the NS1 gene. There is currently not much knowledge about the functions of these genes.

2.1.2.2. The right ORF

The right ORF located between nucleotides 2241-4346 (NC001662) encodes the viral capsid proteins (VP): VP1 and VP2. The VP genes are important for determining host range and virulence. For example, the N-terminus of VP2, amino acid 1-220, has been suggested to play a role in AMDV host range and for its ability to grow in cell culture (Bloom et al. 1998), and the VP2 amino acid 420 has been proposed to increase viral fitness by prevention of caspase cleavage (Cheng et al. 2010). VP2 amino acid 428-446 functions as a small part of the capsid, which might have importance for immune pathogenesis by defining AMDV host range (McKenna et al. 1999). But whether or not this change

also results in increased pathogenicity is unclear (Sang et al. 2012; Oie et al. 1996; Bloom, Alexandersen, et al. 1988). Other VP sites, such as amino acid (aa) 395 and 534 (Bloom et al. 1998), 434 (Oie et al. 1996), have been linked to pathogenicity.

2.1.2.3. Regulatory elements

Promoters are genomic regions where transcription factors are gathered to enable the RNA polymerase to begin transcribe the messenger RNA (mRNA). TATA-boxes are transcription factor binding motifs, and thus compose an important part of the promoter (Buratowski 1994). In AMDV, a promoter that initiates transcription of all mRNA have previously been identified at nucleotide 151-160 (P3: GTATATAAGC) (Bloom, Alexandersen, et al. 1988), in addition to a plausible promoter P36 around nucleotide 1744 (Qiu et al. 2006; Bloom, Alexandersen, et al. 1988). Eight TATA-boxes have been suggested; two confirmed functional at nucleotide 154 (TATAA) and 1729 (TATTAA), both in the left ORF, and six additional boxes at nucleotides 665, 818, 2546, 4136, 4394, and 4468 (AATAAA) (Bloom, Alexandersen, et al. 1988). Internal polyadenylation sites (pA)p's at nucleotide 2561 and 4391 have a proposed major role in AMDV replication (Huang et al. 2012). Studies in the related parvovirus Minute Virus of Mice (MVM) have showed that the NS1 GKRN-region contains a NS1 recognition site at amino acids 337 to 344 (ACCAACCA), which together with an upstream nicking site initiates viral replication in that virus (Christensen et al. 1997).

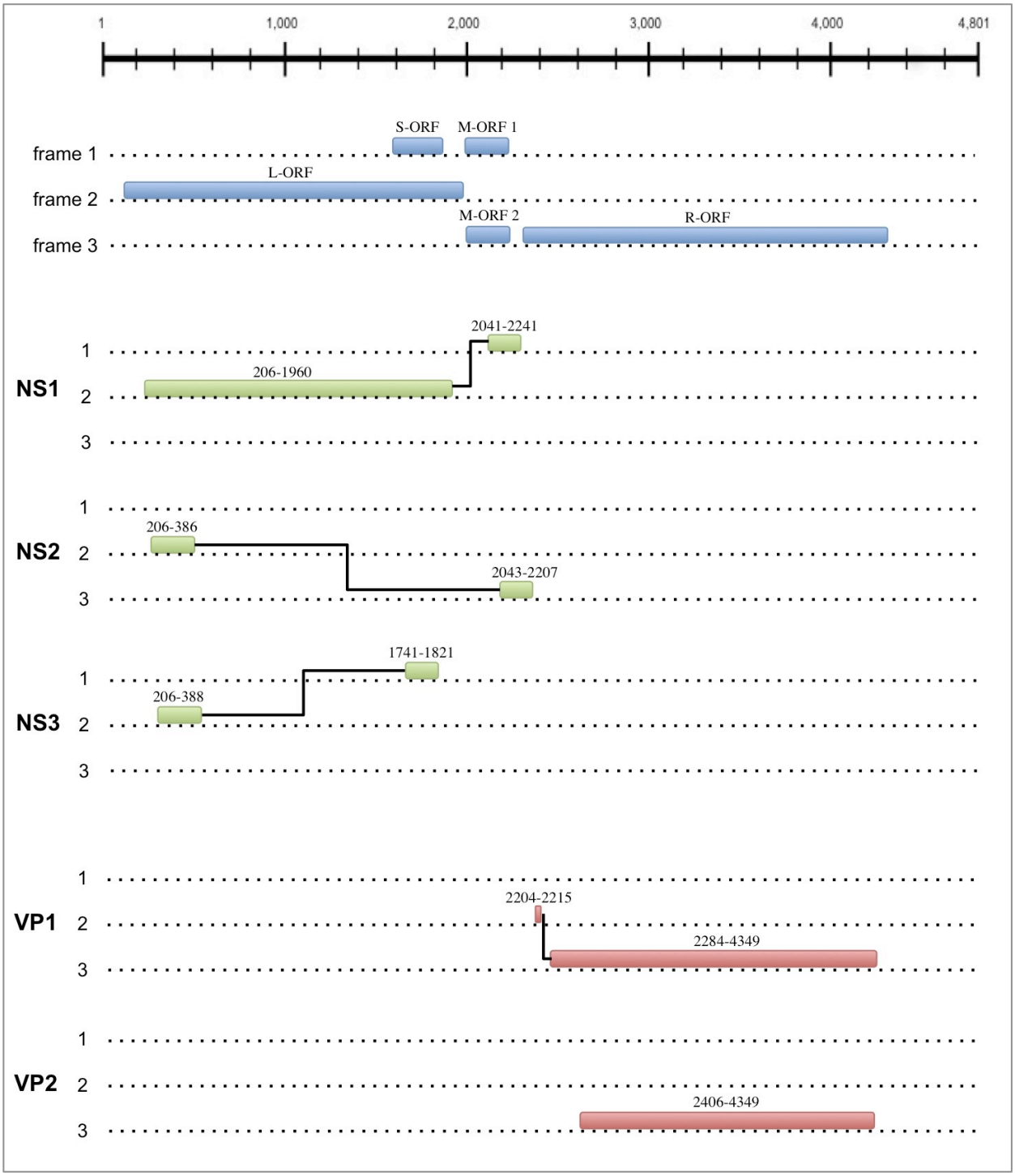


Figure 6. Genomic organisation of the AMDV genome. Illustration of the open reading frames (ORF's) and splicing of the major proteins. The nucleotide positions refer to the reference AMDV-G genome (NC001662). Nucleotide positions: 116-1975 left ORF (L-ORF), 1535-1825 small ORF (S-ORF), 1993-2209 mid ORF 1 (M-ORF 1), 1983-2204 mid ORF 2 (M-ORF 2), and 2241-4346 right ORF (R-ORF). From Manuscript 1 (Hagberg et al. 2016).

2.2. EVOLUTIONARY MECHANISMS

2.2.1. Mutations and substitutions

The central dogma in molecular biology is that DNA is translated into messenger RNA (mRNA), which in turn is transcribed into proteins. The proteins are made up of amino acids encoded by triplets of DNA nucleotides (codons). Depending on the starting codon, a given stretch of DNA can code for three different mRNA's, and is therefore said to have three different reading frames. In order to preserve the essential functions of a genome there are mechanisms acting to protect its DNA from becoming altered during the process of replication. The DNA polymerase in eukaryotic cells has an efficient proof-reading function correcting for base miss-matches with a fidelity, or error-rate, in the range of $1x10^{-7}$-$1x10^{-8}$ subs/nucleotide (Maclachlan et al. 2011), and there is the fact that most amino acids can be encoded by more than one codon. However, there will inevitably be failures in the protection system, which may result in substitutions of a nucleotide, also called a single nucleotide polymorphism (SNP) or single nucleotide variant (SNV).

Mutation rate is the rate by which mutations are generated in a genome, and it can be defined as the number of genetic errors accumulated either per unit of time, per generation, or per round of genomic replication. Known sources of DNA mutations are e.g. DNA polymerase reading errors, base modifications by other cellular enzymes, or introduction of insertions and deletions during replication and recombination (Maclachlan et al. 2011). *Substitution rate*, on the other hand, is the rate by which the induced mutations are fixed at population level, and is denoted as the number of fixed mutational changes per nucleotide site per unit of time (often in years). This rate is a product of the underlying mutation rate, the generation time, the effective population size, and fitness (where advantageous mutations are fixed at a faster rate than at neutral or negative mutation pressure). The effective population size (E_n) is defined as the size of a population with random mating and the same gene frequency changes as in the population under study, and is of importance for the substitution rate (Melorose et al. 2009).

Drake (Drake 1993) suggested that there in viruses is a relationship between mutation rate and genome size resulting in a "universal rate" of $3.4x10^{-3}$ mutations per genome per genomic replication. This assumption however is simplified, as e.g. double-stranded DNA viruses are generally more stable, and utilize the fact that their larger genomes can encode for factors that affects the host immune system or induce mechanisms that protects the virus (Maclachlan et al. 2011). Smaller viruses, and especially RNA viruses, with limited genomic sizes, are hence more dependent on the host cell, and instead use approaches such as mutations in order to avoid the immune defence. Most RNA viruses

have substitution rates in the magnitude of $1x10^{-3}$ subs/site/year (Duffy et al. 2008), which is much higher than e.g. the DNA-polymerase in eukaryotic cells (table 1). The high substitution rates in ssDNA viruses is believed to be due to un-methylated genomes, double stranded intermediates inaccessible to host enzymes, or perhaps due to viral proteins interacting with host factors thereby altering the polymerase (Duffy et al. 2008). A *seemingly* high mutation rate could be the effect of recombination and frequent positive selection.

	viral species	genome	size (nt)	substitution rate (subs/site/year)	reference
AMDV	*Parvoviridae*	ssDNA	~4.800	$1.5x10^{-3}$-$5.7x10^{-4}$	(Hagberg, E.E., Larsen, L.E., Krarup, A., Pedersen, 2017)
Canine parvovirus (CPV-2)	*Parvoviridae*	ssDNA	~5.000	$4x10^{-4}$	(Shackelton, Parrish, Truyen, & Holmes, 2005)
Porcine parvovirus (PPV)	*Parvoviridae*	ssDNA	~5.000	$3x10^{-4}$ - $5x10^{-4}$	(Streck, Canal, & Truyen, 2015)
Human parvovirus B19	*Parvoviridae*	ssDNA	~5.500	$1x10^{-4}$	(Shackelton & Holmes, 2006)
Canine parvovirus (general)	*Parvoviridae*	ssDNA	~5.000	1.09-$1.78x10^{-4}$	(Allison et al., 2013)
FMDV	Apthovirus, *Picornoviridae*	ssRNA	~8.500	$8.3x10^{-3}$	(Hanada, Suzuki, & Gojobori, 2004)
HIV-1	Retroviridae	dsRNA	~9.180	$2.4x10^{-3}$	(Hanada et al., 2004)
Eukaryotic DNA polymerase				$1x10^{-7}$-$1x10^{-8}$ (mutation rate)	(Maclachlan, Dubovi, & Fenner, 2011)

Table 1. Overview of selected viruses and their respective substitution rate, type of genome, genome size, and reference.

There are different types of mutations: transitions, which is the exchange within the purine base group (A, G) or within the pyrimidine base group (T, C), and transversions, which is the exchange *between* the two groups (e.g. purines to pyrimidine's, and vice versa) (Melorose et al. 2009). Due to chemical properties, transitions occur more frequently than transversions. Insertion and deletions are mutation events where one or several nucleotides are inserted or deleted from the genome and can more drastically impact the reading frames.

Point mutations affecting protein binding sites or translation initiation sites potentially have more impact than mutations in other genomic regions due to their impact on the replication cycle. As the

first position in a codon is often the main determinant for translation, mutations at this position more often result in the production of a different amino acid compared to point mutations in the second or third position. The redundancy in the genetic code allows the same amino acid to be translated from different combinations of nucleotides. Therefore, mutations that do not affect the encoded amino acid can be called synonymous, and mutations that result in a change of amino acid non-synonymous. Synonymous changes and the changes just altering the amino acid into one with similar properties is the most common, and have little if any impact on the encoded protein. Amino acid changes can also be grouped depending on if and/or how much it alters the function of the downstream protein. Some changes produce amino acids with similar polarity and will not affect the protein, while others result in a very different or non-functional protein.

The ratio between non-synonymous and synonymous amino acid changes is a common tool for estimating the selective pressure acting on a sample of viral sequences (Melorose et al. 2009; Fields et al. 2007). Selection is said to be neutral or nearly neutral when there is a simple relationship between mutation rate and substitution rate, and indicates mutations are not impacting the fitness in the environment from which the samples came. When the number of synonymous substitutions is larger than the non-synonymous, the ratio between them becomes less than one, and the selection pressure is said to be *negative or purifying*. This is most profound in viruses well adapted to their environments, as a large proportion of mutations, especially those in key regulatory elements, would be deleterious in that particular environment (Fields et al. 2007). *Positive* selection on the other hand, favour viral variants with a survival advantage under the selective constraints in that particular environment.

2.2.2. Genetic recombination and quasispecies

Recombination events can occur when two or more related viral strains simultaneously affect the same cell or host (Maclachlan et al. 2011). The process is somewhat different in DNA and RNA viruses. Intermolecular recombination refers to an enzyme-mediated exchange of nucleotide segments between viral strains during replication and is the most common means of recombination in dsDNA viruses and in some RNA viruses. This process sometimes involve host gene elements, which can result in better avoidance of the hosts immune system e.g. during persistent infections (Posada et al. 2002). Homologous recombination occurs when nucleotide sequences from the same genomic coordinates are exchanged between organisms (Melorose et al. 2009). Reassortment on the other hand, mainly occurs in RNA viruses, and involves exchange of genome segments between strains resulting in a viable and stable new version of the virus. Recombination events are important for viral

evolution as they can contribute to more disruptive changes than e.g. point mutations and the generation of quasispecies can. Recombination can expand host range or increase virulence (Melorose et al. 2009). It has been shown that multiple parvoviruses can infect the same host, which possibly explains the relatively high recombination rates in parvoviruses compared to other DNA viruses (Shackelton et al. 2007).

The concept of quasispecies is based on that each viral species is defined by some characteristics (phenotypic and genotypic) from which most individual genomes differ to some extent, thereby creating a "pool" of possible variants of the virus. The adaptive potential of a virus can be increased by this mutant spectrum, as different phenotypes already are present in the population or can be generated quickly (Fields et al. 2007). Quasispecies are most commonly described in quickly evolving viruses such as HIV and RNA viruses, and the concept should be kept in mind in relation to e.g. viral outbreak/epidemic investigations, as the consensus sequence for each sample often used for downstream analyses represents an "average nucleotide sequence" which might not exist in the real population.

2.3. DIAGNOSTIC METHODS FOR AMDV

2.3.1. Clinical diagnostics

The veterinarian often performs the first line of diagnostics at a population level on the farm. Clinical signs are mainly non-specific (Jensen et al. 2015) and reflected in production parameters such as reduced growth, fur quality and fertility, or increased mortality. In well-managed farms these are often the only signs of disease, however in more advanced disease cases there can additionally be sprinkled white hairs, bulldog noses, and un-thriftiness. In kits, differential diagnoses to AMDV can be pneumonia due to influenza or bacterial infections (Maclachlan et al. 2011). Post mortem analyses aims to rule out other disease, as the macroscopic pathological signs of AMDV are vague, if present at all. Macroscopic lesions can include splenomegaly, renal swelling or petechiae, and enlargement of mesenteric lymph nodes. Histologic lesions involve plasma cell infiltration in the spleen, lymph nodes, bone marrow, kidneys, liver, as well as interstitial pneumonia or fibrinoid arthritis (Jensen et al. 2015).

2.3.2. Serological diagnostics

The serological diagnostic methods most often detect the immune responses toward the virus and not the virus per see. The first method used for field diagnostics of AMDV was the iodine agglutination

test, and in the early 1970's the counter current immune electrophoresis (CIEP) was subsequently developed for diagnostics of AMDV (Cho & Ingram 1972). This assay detects immunoglobulin (IgG and IgM) in serum samples loaded into a well on the cathode side of an agarose gel. On the opposite side viral antigen is loaded, and after running the gel positive samples will create a visible antibody-antigen complex seen as a line between the wells. The antigen in this case is the entire AMDV particle, and binding is presumed to be directed both to VP and NS epitopes (Bloom et al. 1982). The method is however labour intensive and requires trained personnel, and it has been shown that the specificity of the test is affected by the manual readout (Dam-Tuxen et al. 2014).

Enzyme-linked immunosorbent assays (ELISA) allows for high-throughput screening of antibodies in blood samples. A fully automated ELISA was implemented in Kopenhagen Fur in 2015 (Dam-Tuxen et al. 2014). Briefly, the farmer takes out a small blood sample and puts it on a filter paper attached to a plastic frame (dry blood spot card). In the laboratory, the sample is punched from the card and treated with an extraction buffer. After extraction the sample is transferred to a microtiterplate coated with viral antigen containing whole AMDV-G particles to which the AMDV antibodies (if present in the sample) bind. After incubation a secondary antibody labelled with horseradish peroxidase (HRP) is added to the well and it will bind to any attached anti-AMDV-G antibodies. Finally, a fluorescent substrate for HRP is added and the antibody titre can be quantified. ELISA allows for high throughput screening and is suitable for automation, and sampling can be done on easy to ship and store dry blood spot cards. However, due to cross reactivity with the labelled secondary antibodies and insufficient blocking of unspecific binding sites of the coating antigen indirect ELISA is, as other diagnostic tests, not 100% specific.

The main limitation with the serological methods is in regards to detection of early disease, where the circulating antibody levels are either absent or below the detection limit (Jensen et al. 2015). Additionally, wild type AMDV cannot be grown in cell culture and therefore diagnostics rely on either detection of the virus using molecular based methods or the antibodies produced toward it. Thus, in order to diagnose AMDV during early disease stages molecular diagnostics could provide a valuable tool.

2.3.3. DNA based diagnostics

Nucleic acid based diagnostics utilises the polymerase chain reaction (PCR) to detect the viral DNA, but does not imply that there is viable infectious virus in the samples. Positive PCR results can further be verified by sequencing and the data can, in addition to detection, be used to subdivide and classify

the pathogen of interest, and to create a knowledgebase for downstream analyses such as phylogenetics (chapter 2.4). Since AMDV replicates in dividing cells of the lymphatic system, the spleen and lymph nodes, which harbour these cells, are the most commonly used tissues for diagnostics and extraction of nucleic acids. The viral DNA is isolated from the sample material through a purification step, often performed using a commercial kit, which removes potential downstream inhibitors such as acids, cell debris, and proteins. The basic concept behind the purification procedure is to break down the viral capsid to release the DNA, capture the DNA by binding it to a specific column, and to separate it from cell debris and other inhibitors through rinsing. The DNA is eluted in a pure form either in distilled water or a storage buffer ready for further processing.

2.3.3.1. Conventional endpoint-PCR

The polymerase chain reaction (PCR) was developed in 1983 (Mullis et al. 1986) and has become one of the most widely used diagnostic tools specifically detecting target nucleic acids. PCR utilise thermal cycling to create optimal conditions for a commercially engineered DNA-polymerase to extend a given stretch of DNA through the incorporation of artificial/added nucleotides. Short (20-25 bases) nucleotide fragments (PCR primers) are designed to specifically bind a given region in the target DNA sequence at a given temperature. These primers are added to the PCR reaction consisting of an equimolar mixture of DNA-nucleotides, a DNA-polymerase, a buffer, and the target DNA. The reaction is incubated in a thermal cycler with pre-set temperature cycles often including an initial activation of the DNA-polymerase together with denaturation of the target DNA, followed by a number of cycles (often in the range of 35-40 cycles) consisting of denaturation, annealing, and strand elongation/extension. DNA amplification is initiated during the step where the DNA-polymerase binds to the primers. Successful amplification is confirmed by running the PCR products on an agarose gel stained with e.g. ethidiumbromide that makes DNA fluoresce upon exposure to UV-light. The size of the DNA fragment is determined by visual comparison to a DNA ladder.

An endpoint PCR assay targeting a 374 bp fragment of the AMDV NS1 gene was developed in 2011 at the Danish National Veterinary Institute (NVI) (Jensen et al. 2011). It has since routinely been used for plasmacytosis diagnostics followed by a confirmatory Sanger sequencing to increase the specificity, thus the bulk of genetic data available in Denmark arises from this partial NS1 gene region (Jensen et al. 2011; Ryt-Hansen, Hagberg, et al. 2017).

2.3.3.2 Real-time PCR

Real-time PCR elaborates on the concept of thermal cycling as in endpoint PCR, but with the addition of real-time monitoring of the accumulation of fluorescent signal (Higuchi et al. 1993). This removes the labour intensive and potentially hazardous step of agarose gel analysis, thereby decreasing analysis time, reducing risk of contamination, and providing more objective cut-off values. This method also provides better accuracy than endpoint PCR, as the quantification takes place during the exponential phase instead of at the plateau-phase, where in the latter some of the reagents might be used up unequally in the different samples (Design & Optimization 2010). The two most frequently used real-time PCR chemistries are double stranded intercalating chemistry and dual-labelled hydrolysis probes.

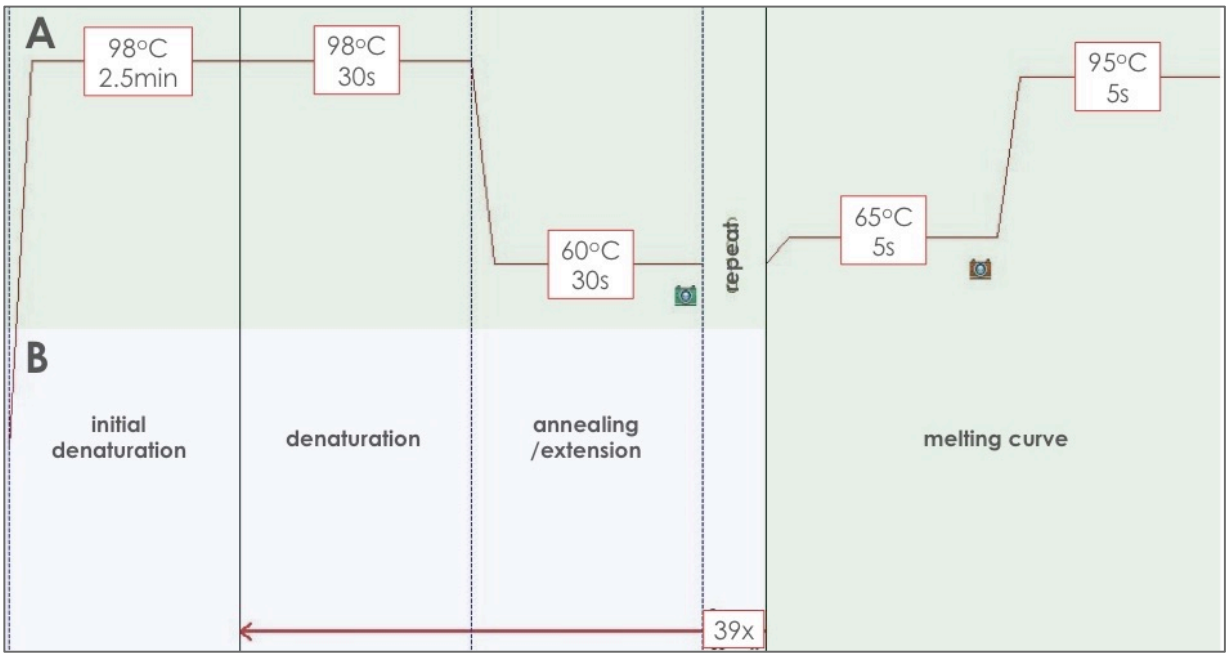


Figure 7. Real-time PCR cycling. Schematic illustration of the cycling during SYBR (A) and hydrolysis probe (B) protocols. Both begin with an initial denaturation, followed by rounds of cycling between denaturation and a combined annealing/extension step. Protocols with SYBR-green can end with a melting-curve.

A sensitive diagnostic PCR assay, whether based on endpoint or real-time technology, is designed to bind a conserved genomic region to prevent mismatch of primers/probes and the target DNA, and to increase the probability of amplification and hence detection of the pathogen. Advantages with the hydrolysis probe chemistry is its increased specificity and that the technology can be multiplexed by incorporating several probes into the same assay to simultaneously detect e.g. several genotypes of a pathogen in a single sample (Dowgier et al. 2016).

2.3.4. DNA sequencing – first generation

The first commercially applied sequencing method with a reasonable throughput was the so-called Sanger sequencing (Sanger et al. 1977). It was developed in the 1980's and has since been modified slightly to fit modern requirements in regards to throughput. Sanger sequencing is sometimes referred to as chain-termination sequencing or dye termination. The input material is DNA, either pre-amplified with PCR or directly extracted from a sample. The former approach is commonly used when there is e.g. low amounts of target DNA or large amounts of interfering DNA in the samples. In short, the sequencing reaction contains the target DNA and is added primers and DNA nucleotides labelled with a fluorescent dye stopping the synthesis upon incorporation into the growing DNA strand. The resulting pool of DNA-fragments (reads) will have different lengths and are used for determining the order of the nucleotides in the DNA-strand, either by running the products on a gel or by registering the fluorescent signals in the sequencing machine. Stretches of up to around 1000 bases can usually be sequenced with reasonable reliability, and the sample preparation is simple for a person with laboratory training. The sequencing step itself is relatively cheap and easy to perform, which has contributed to its popularity. The error-rates however are often high in the beginning and end of the reads, and the output is a single consensus read per sequence with its associated errors (Metzker 2010).

2.3.5. DNA sequencing – next generation

Next generation sequencing (NGS) is a powerful set of technologies that since its commercial introduction in the 1990's/2000's has experienced a sharp decline in the per base sequencing price and an upswing in availability. Today NGS has gained widespread usage in molecular biology, food safety, healthcare, biotechnology, pharmaceutical research and development, veterinary medicine, and human public health. NGS has been successfully applied to characterise entire genomes of important viruses and the genetic information obtained has been used to improve preventative measures (Escobar-Gutiérrez et al. 2012; Jakhesara et al. 2014; Kvisgaard et al. 2013).

The input material for NGS is as for conventional sequencing DNA. The first step during sample preparation is the construction of a so-called library, where the target DNA in each sample is fragmented, either enzymatically or physically, and labelled with specific adapter sequences that allows to distinguish the samples from each other. The latter is important as often many samples are run simultaneously. The library is then PCR amplified (with some variation depending on the NGS technology) to generate a clonal pool of templates ready for sequencing. The basic principle is that reactions containing these libraries sequentially are added DNA nucleotides in a known order, and

when the DNA polymerase incorporates the nucleotides either a light signal (Illumina) or a current (Ion Torrent PGM and Roche 454) is released and registered by the sequencer. The machine sequences (i.e. reads) each of these clonally amplified templates simultaneously (hence the name massive parallel sequencing), generating a large pool of output referred to as the *sequencing reads*. During the data-analysis, the individual reads are assembled based on their overlaps, i.e. put back into a meaningful order. This generates the so-called read-depth, which is one of the major advantages with NGS compared to Sanger sequencing, where one read corresponds to the average of several molecules. Despite the higher error-rates associated with NGS, the read-depth compensates for this as each base and its neighbours are registered several times, thus providing a statistical framework from which a "consensus" base can be defined (Quail et al. 2012; Metzker 2010).

Some of the commercially most frequently used platforms are Illumina® HiSeq and MiSeq (Illumina, Inc., San Diego, CA), Ion Torrent PGM™ (Life Technologies, Carlsbad, CA), SOLiD™ (Applied Biosystems, Foster City, CA), and the 454 Sequencing technologies (Roche, Basel, CH). They differ slightly in regards to the depth, length, and quality of the reads they generate, which has been reviewed in more detail elsewhere (Quail et al. 2012; Metzker 2010). Illumina® has for a long time dominated the market with their MiSeq and HiSeq machines, which produce reads in the somewhat shorter range (40-300bp), but have other advantages, such as paired-ends that facilitates mapping and *de novo* assembly. The Ion Torrent PGM™ and the 454 Sequencing technologies are known for having problems differentiating between long stretches of homopolymeric regions, due to how synthesis is registered (Quail et al. 2012). Briefly, the sequencing is based on registration of a current created from the release of protons upon incorporation of nucleotides into a growing DNA strand. The machine sometimes cannot accurately distinguish between the signals generated by e.g. five or six identical bases in a row. However, these technologies provide other advantages such as being easier and cheaper to run, and by having longer average read-lengths (150-400bp). The latter is useful for assembling longer sequences, as fewer reads and overlaps will be required. There are other more novel technologies, sometimes referred to as the third generation of sequencing, which does not rely on a PCR step for pre-amplification and provides longer reads. However still associated with higher error-rates.

2.3.5.1. NGS data analysis

The two most commonly used approaches for processing NGS data is either in a graphical user interface provided by software-packages such as Geneious (Kearse et al. 2012) or CLC Genomic Workbench (Anon n.d.), or to combine individual tools into customized pipelines that can be

executed by running scripts in the command line, often in an Unix-environment. The latter allows for incorporation of up-to date versions of each tool, precise control over the applied parameters, excellent reproducibility, and batch-analyses. However, it puts higher requirements on the operator's skills and experience. The two most common approaches for assembling the reads are to gather them using a guidance reference (mapping) or to assemble them *de novo*. Mapping is commonly used when there is a reference genome available, and is often a good starting-point due to its relative ease and lower computing power requirements. However, *de novo* assembly potentially provides a more unbiased and realistic sequence, as it is not influenced by a reference, but is on the other hand sensitive to low quality reads and uneven coverage. The extra layer of information from having a high read-depth can be used e.g. to investigate nucleotide variation at individual sites, look for quasispecies and haplotypes, or to generate a consensus sequence to use as input for phylogenetics, which is described in more detail below. In relation to viral outbreak investigation, the most common workflow is to use a consensus DNA-sequence from each sample (Gilchrist et al. 2015). The consensus sequences can be stored in fasta-format, a plain text file allowing for easy and convenient manipulations, taking up much less space than the raw-data. Regardless of data-processing preferences the main steps are similar and can be divided into pre-processing, assembly, and post-processing. These steps are described in more detail in Chapter 3 - materials and methods.

2.4. PHYLOGNETIC ANALYSES

Phylogenetic analyses aims to reconstruct the past (i.e. the history) of a given species (or set of samples or taxa) and put it on an evolutionary scale (e.g. years or substitutions per site per year) by using input information such as morphological data or a multiple sequence alignment (MSA). NGS and especially data from whole genome sequencing allows for numerous possibilities in regards to phylogenetic analyses, e.g. the investigation of relationships between samples at multiple levels; intra-individual, inter-individual, within a population, or between populations. Figure 8 summarises some useful terms to have in place in relation to phylogenetics.

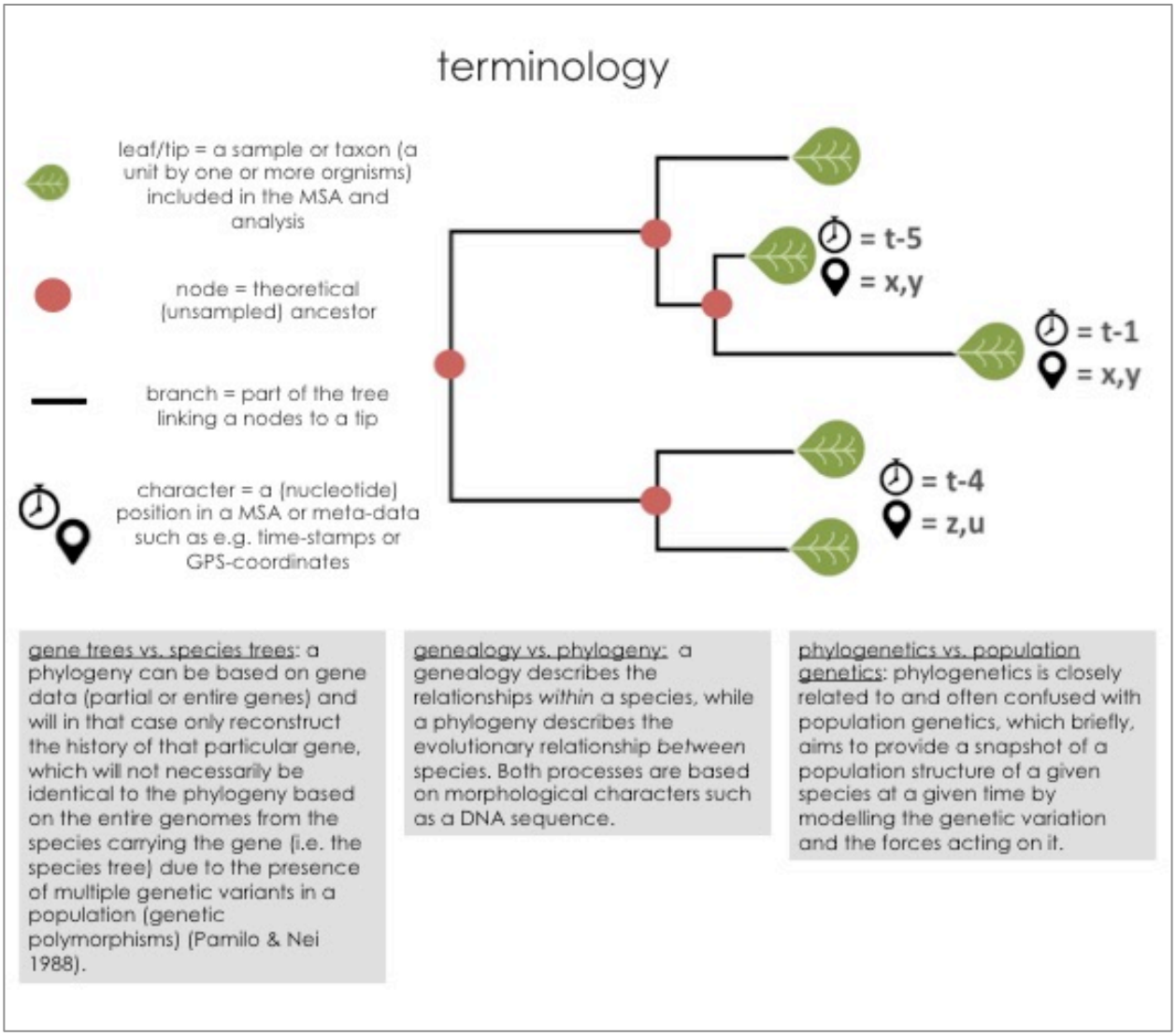


Figure 8. Selected terminology. Summary of some useful terms to have in place in relation to phylogenetics.

2.4.1. Phylodynamics

The term 'phylodynamics' was coined in 2004 (Grenfell et al. 2004) and represents a somewhat new field, aiming to put genetic data (phylogenetics) on a calendar time-scale, e.g. to extract the signature of genetic data originating from infected hosts in an epidemic and to quantify population dynamics such as epidemic spread. Key components when applying phylodynamics in outbreak investigations are to determine whether the case strains are the same as in the rest of the population, and to discriminate between background noise (Quick et al. 2015). During a hospital outbreak of Salmonella, the importance of relating the whole genome sequences into a community and a country context was highlighted (Quick et al. 2015), and others have also illustrated the strengths of whole genome sequencing and the need for integration with (epidemiological) metadata (Snitkin et al. 2012).

Traditional epidemiology aims at estimating parameters such as distributions of incubation and infectious periods, rates of clearance from the population, and basic and effective reproductive numbers (du Plessis & Stadler 2015). These estimations are based on surveillance data, e.g. clinical and demographic data such as incidences from case reports and hospital records. This information is used for creating a transmission tree representing the relationships between the sampled hosts. The disadvantages with such approach are that surveillance data is susceptible to human errors: especially in times of stress, in geographic regions with poor infrastructure, and in the absence of clear guidelines, and it ignores the pathogens genome - a potentially important trait (du Plessis & Stadler 2015).

Epidemiological transmission trees represent the relationship between the sampled individuals, while phylodynamics additionally take into account only sampling a proportion of the infected individuals (Grenfell et al. 2004). However, since the transmission events must take place after the lineage-splitting event (which can be quite long time for quickly evolving), these trees will be different (fig. 9).

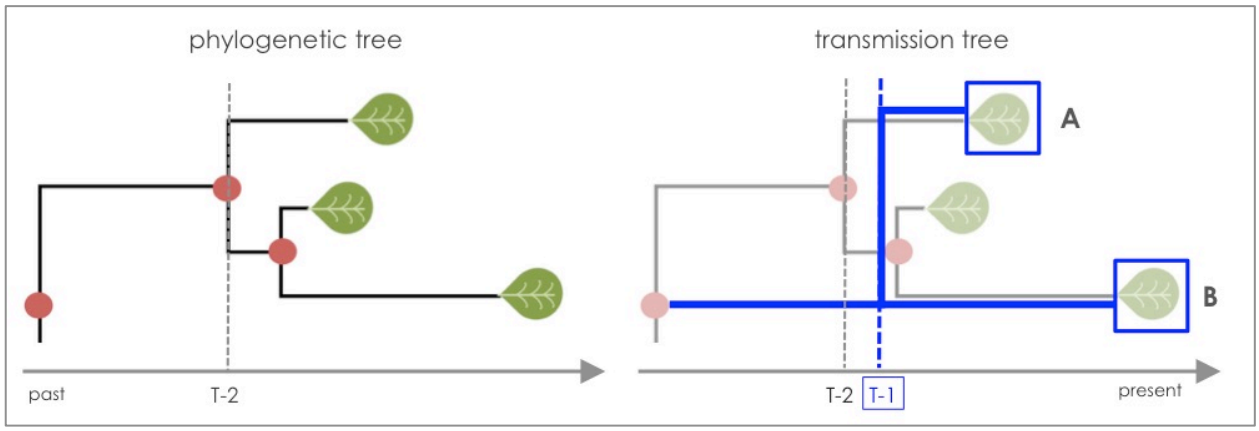


Figure 9. The relationship between a phylogenetic tree and a transmission tree. The phylogenetic tree shows the relationship between the tips and their evolutionary most recent common ancestor (T-2). The transmission tree (blue overlay) illustrates the relationship between tip A and B based on the epidemiologically known transmission event (T-1), which in this case occurred *after* the lineages latest evolutionary split (T-2).

2.4.1.1. Epidemiological aspects of phylodynamics

The most common approach for describing infectious disease dynamics is to fit so called compartmental models to surveillance data. For a given set of parameters these models describe the temporal dynamics in subpopulations characterised by their disease status. One of the frequently used models is the Susceptible-Infected-Recovered (SIR) model, in which the number of susceptible

infected, and recovered (and immune) individuals are described. The perhaps most important statistics in epidemiology are the basic (R_o) and effective (R_e) reproductive numbers. The R_o is the average number of secondary infections caused by an infected individual over the course of an epidemic (given a completely susceptible population). An $R_o>1$ equals an on-going epidemic and that the number of infected will increase. The R_e is the average number of secondary infections caused by an infected individual at some later time during the epidemic, and is used for describing dynamics later in an epidemic or if there is pre-existing immunity in the population. The two common approaches for quantifying epidemiological dynamics are: 1) by fitting a demographic model to the genealogy thereby getting a proxy for the transmission chain between the sampled individuals, or 2) direct quantification by inferring both the genealogy and epidemiological dynamics from sequencing data (i.e. phylodynamics). This thesis will only focus on the latter.

The perhaps most straightforward approach for finding the origin of an epidemic is to determine the non-epidemic genotype most closely related to the epidemic, i.e., the molecular sequence clustered most closely to the epidemic strain on a phylogenetic tree. This is intuitive but heavily dependent on the collected data (Kühnert et al. 2011), and if one fails to identify a strain sufficiently closely related to the epidemic strain the phylogenetic tree cannot provide the origin. Interspersion of an emergent viral strain within other strains in a phylogenetic tree is often interpreted as a sign of multiple independent introductions. However, due to incomplete taxon sampling resulting in undercounting these events, the number of clusters does not equal the number of transmission-events. Un-sampled sequences might exist and could split the clusters, or there might be additional un-sampled clusters (Kühnert et al. 2011).

2.4.1.2. Divergence time estimation and dating phylogenies

An alignment of sequences sampled at different points in time comprise a heterochronous dataset, which allows for additional analysis options compared to if the dataset was contemporaneous, e.g. it can be used to estimate substitution rates and to calibrate divergence times to a calendar scale (Kühnert et al. 2011). The branch-lengths in such phylogenetic tree will therefore represent a unit of time (e.g. years), and not genetic distance as in distance based trees.

Calibration of a phylogenetic tree to a time scale can be done either on the trees internal nodes or on the tips/leaves (the samples). The former approach is common in e.g. ecology and evolution, where exact sampling dates are often not available but events such the known time of extinction of a species are and can therefore be used to calibrate the time of an internal node in the tree. Working with

viruses one operates on a totally different time-scale (often days to decades), and during outbreaks/epidemics it is common to register various epidemiological meta-data such as sampling-dates and locations, which can be used as input for other analyses. And as mentioned above, transmission events are often more recent than the most recent common ancestor (MRCA) of the sampled sequences, and thus the sampled epidemiological transmission tree is not necessarily the same as the reconstructed phylogeny (fig. 9).

2.4.2. Phylogenetic methods

This chapter aims to provide an overview of the theoretical basis for the phylogenetic methods applied in this project.

2.4.2.1. Distance based methods

The most straightforward way to evaluate the relationships between a set of sequences is to evaluate the pairwise distance between them. In a given alignment the genetic distance between any pair of sequences, i.e. the number of mutations separating them, can be counted and entered into a distance-matrix (observed distances). The aim is to create a tree where these observed distances approximate the patristic distances (a measure created by summing the branch-lengths between two tips). The "best" tree is found by comparing topologies and selecting the one generating the smallest overall deviation between observed and patristic distances (cluster analysis), or the smallest sum of branch-lengths (minimum evolution). Potential drawbacks are that the overall rate of evolution across the tree is assumed to be the same, and therefore these methods can be sensitive to unequal rates between lineages (Melorose et al. 2009), and that the observed distances might not reflect the evolutionary distances since multiple (unobserved) substitutions at the same site can obscure the real distances (Holder & Lewis 2003).

2.4.2.2 Maximum Likelihood methods

Briefly, in maximum likelihood (ML) the goal is to find the tree with the highest likelihood of producing the observed data given the model. The model consists of parameters such as the trees, branch-lengths, nucleotide frequencies, substitution-rates, and clock-rates. E.g. if point mutations are considered as chance events, the probability of finding a mutation a long a branch in a tree can be calculated. ML aims to maximise the probability of observing the sequences in the data set by determining the tree parameters. The difficult task in ML is not to find the branch lengths that maximise the log-likelihood function for a tree, but to search all possible tree topologies and find the tree that maximises the overall likelihood in the dataset (Melorose et al. 2009). ML is due its

underlying calculations computationally demanding, and the probability of getting the observed data is not related to the probability that the underlying model is correct (Melorose et al. 2009).

2.4.2.3. Bayesian phylogenetic methods

The overall aim with the Bayesian phylogenetic approach is to develop a hypothesis/model for the evolutionary process, and to find the tree with highest probability given the data. In similarity with ML the Bayesian methods also attempt to account for unobserved character changes in addition to quantifying the degree of belief in the results. The main difference between Bayesian and traditional (frequentist) statistics is the definition of the term *probability*. In traditional statistics, a probability is considered the long run frequency of an event in a repeatable experiment, i.e. if you toss a coin enough times you should expect a probability of 0.5 of getting a head. The belief is that data is used only to estimate the real/true population parameters. However, Bayesian statistics consider probability as a quantification of the degree of belief, that is; how much do we believe in the outcome of the results *before* looking at the data? This so-called prior belief is used for obtaining/calculating a posterior probability estimate for how much we believe in this parameter *after* looking at the data. Prior to performing a Bayesian phylogenetic analysis, a model needs to be described for the data. The model parameters (e.g. the phylogenetic tree) are treated as random variables and their uncertainties are specified by probability distributions (the priors). These prior distributions and their uncertainties are then updated based on sampling of the data, and assessed by their posterior probabilities (Huelsenbeck & Rannala 2004). The posterior probability of a tree is conditioned on the data and the model being correct.

The probability of a particular value of a parameter, given a set of data, can be described with the posterior probability distribution, also referred to as Bayes' theorem (fig. 10). Basically, the knowledge about a hypothesis H is updated based on the data D, and the equation can be solved by rearrangement and input of the known parameters thereby estimating the unknown ones. Since most of each posterior probability is likely to be concentrated in a smaller part of an often large parameter space, it is virtually impossible to derive this distribution analytically (Melorose et al. 2009). However, the parameters can be sampled using Markov Chain Monte Carlo (MCMC) integration, and summarised into estimates of the posterior probability distributions of each parameter (Drummond et al. 2002). The Markov chains will, if ran long enough, converge to an equilibrium state regardless of where it started.

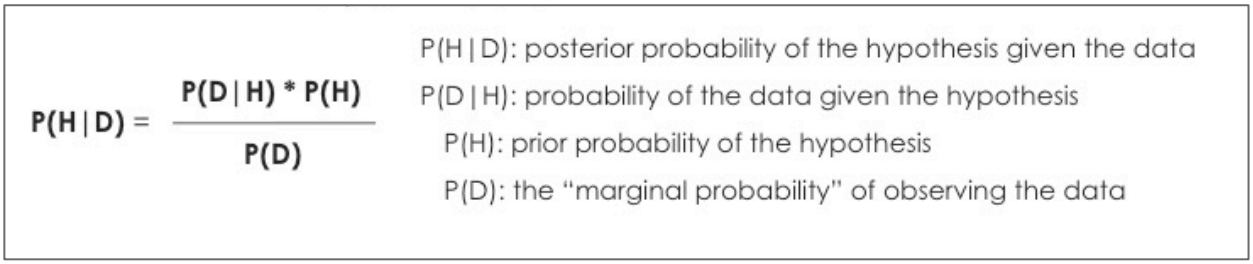


Figure 10. Bayes' theorem. The probability of the hypothesis given the data.

2.4.3. Models of DNA evolution

The underlying theory of DNA evolution and the concept of superimposed substitutions impact the downstream analyses. Briefly this concept means that despite only one nucleotide change is observed, there might have been a range of changes, from e.g. an A in internal node to a C observed in the leaf (fig. 11).

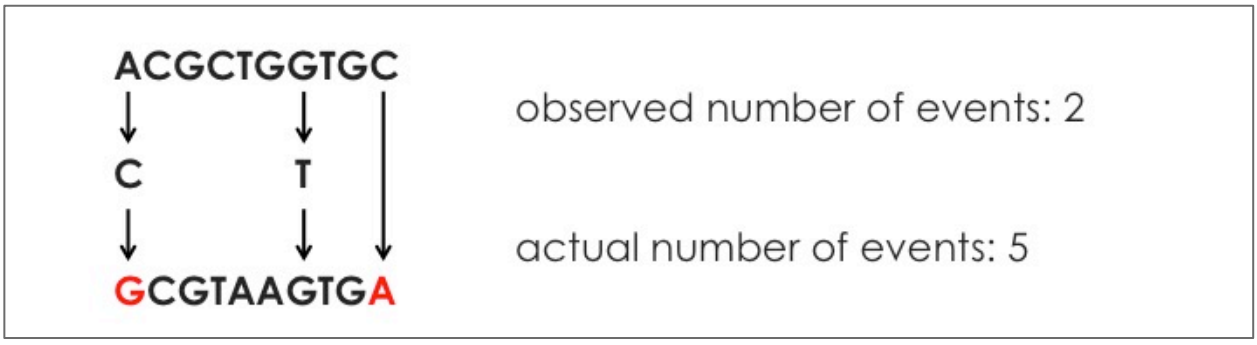


Figure 11. **Superimposed nucleotide substitutions.** An alignment of two sequences where in the first position there has been a shift from A to C to G, representing two distinct evolutionary events, despite that only one change is observed. In the seventh position there was change from a G to a T, and back again to a G, resulting in no observed change.

Given the observed distance (i.e. the sequence alignment) phylogenetics try to infer the real distance. For this purpose, a model is needed, which is considered a stringently phrased hypothesis where the substitution model corrects for unseen mutations in the observed data (the alignment). The simplest model is the Jukes and Cantor (JC), which assumes equal nucleotide frequencies (0.25) and equal substitution rates, meaning that all substitutions are equally likely. The Kimura2-model (K2) (Hasegawa et al. 1985) has two separate substitution rates stating that changes within a group of pyrimidine's or purines (transitions) occurs more frequently than the change between the groups (transversions). In this model the nucleotide frequencies remain equal. In the Felsenstein (F81) model the substitution rates are equal but the nucleotide frequencies can vary. The Hasegawa, Kishino, and Yano (HKY85) model (Hasegawa et al. 1985) distinguishes between transition and transversion rates,

and allows nucleotide frequencies to vary. Finally, the most complex are the General Time Reversible (GTR) models, which assumes the amount of change to be equal in all directions and therefore nucleotide frequencies remain constant. It is more heavily parameterised as every pair of substitution has their own rate.

The models described above can be extended with options regarding the variation across the sites in the alignment. The invariable site model takes into account that certain positions in the genome can be assumed never to vary or to be under very strong selection where every mutation would be removed and therefore never observed. Alternatively, non-variable sites can be considered a variable, meaning that all sites are assumed to mutate at the same rate. Applying a gamma rate distribution equals to the assumption that substitution rates vary at different sites in the genome. Some parts of the genome might undergo more changes than others, and sometimes the rates are the same all over, but due to selection mutations in important coding regions these individuals are removed and will not be observed due to reduced fitness. Some sites might be under strong selection and therefore removed incorrectly appearing as if they do not vary.

Before performing a phylogenetic analysis, it is therefore often advised to estimate the most appropriate nucleotide substitution model for that particular data set. In general, choosing a too complex model (over parameterisation) introduces the risk of over fitting the model to the data, in the worst case assigning each data point a parameter, causing the model to represent of a list of data points (fig. 12). For nested models, adding more parameters always results in a better fit to the data, but not necessarily a better description. The result can be you are fitting the noise instead of capturing the underlying reality. The goal when selecting a DNA nucleotide model is to determine if the more parameterised model fits the data (alignment) significantly better taken into consideration the added number of parameters. There is software available for this purpose, but despite being frequently applied in literature, their usefulness is debated due to differences in selective pressure over the genome, as this might result in different DNA nucleotide substitution models and thus conflicting results (Melorose et al. 2009).

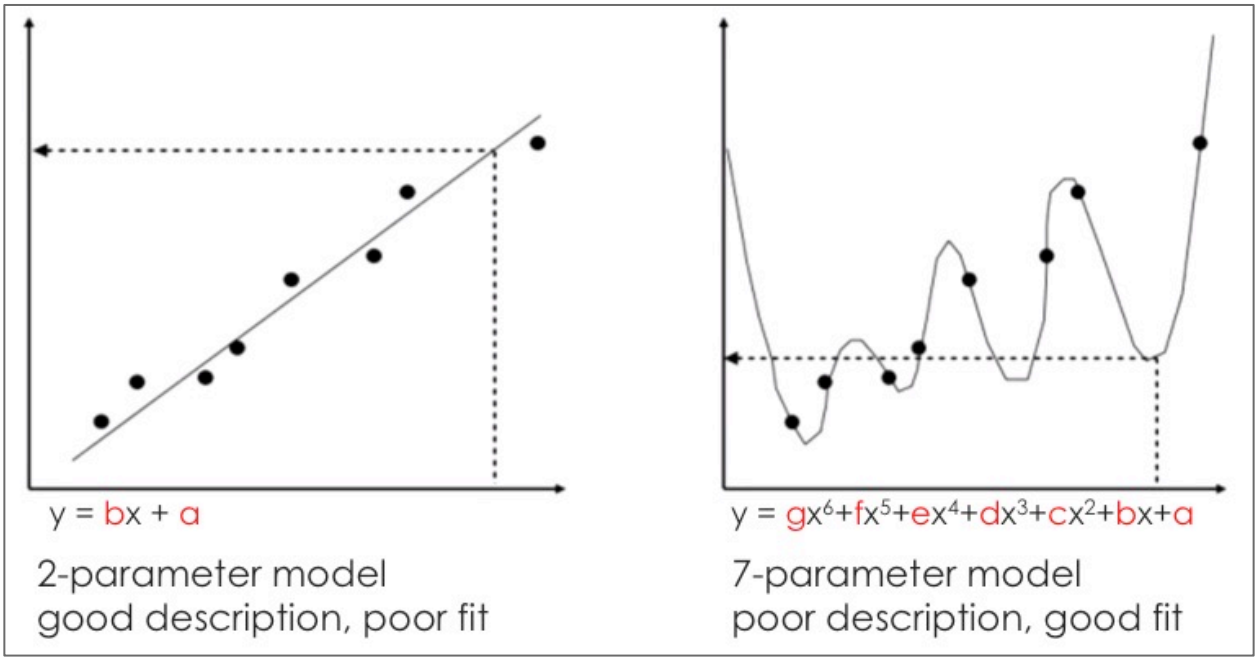


Figure 12. Graphical illustration of over-fitting. The model to the left describes the data well but is a poor fit, while the model to the right is a good fit but with poor description. Modified from Anders Gorm Pedersen's Coursera course "Molecular Evolution" (Gorm Pedersen 2013).

2.4.4. Clock models

The molecular clock is a quantitative description of the rate at which genomes evolve. It explicitly reveals the number of nucleotide changes per unit of time in a particular tree or on a branch, and thus enables the calculation of absolute time-points, such as the age of internal nodes or time-points for transmission events. This is in contrast to non clock-based analyses, where the indicated number of changes per branch is a product of both time and rate, meaning that a long-branch can be the result of a slow rate for a long time, or vice versa.

The simplest model is a strict clock, which assumes the same rate of evolution across all branches in a tree. This assumption is not necessarily the most realistic, but is often sufficient for describing dynamics in intra-species datasets with low rates of variation between branches (Brown & Yang 2011; Drummond et al. 2007). However, more complex datasets often require more comprehensive models to describe the evolution in a realistic manner. Drummond et al. (2006) introduced the so-called relaxed clock models, which allow for rates to vary across the branches in a tree. The mean rate is described either with an exponential or log normal distribution, also referred to as relaxed exponential and relaxed lognormal molecular clocks (Drummond et al. 2006).

In evolutionary phylogenetics fossils are often used to calibrate molecular clocks and to provide information about evolutionary rates. However, in virology observations such as sampling-dates are often more appropriate. Obtaining sequence data from genomes of replicate populations (so called mutation accumulation studies) can be used for calculating mutation rates, however it should be kept in mind that this potentially reflects the substitution rate rather than the mutation rate, as the effects of natural selection cannot be ruled out.

2.4.5. Models of tree evolution

The process by which the tree has evolved is important for its topology, and can be either viewed backwards in time (coalescent theory), or forwards in time (Birth-Death (BD) models). More classic epidemiological approaches such as the susceptible infected recovered (SIR) models can also be incorporated into a phylogenetic framework (briefly described below).

2.4.5.1. Coalescent models

The coalescent framework allows for inference of population history from sequencing data by describing the statistical properties of the genealogy underlying a random sample of individuals to the population from which they were sampled (Kühnert et al. 2011; Rodrigo A & Felsenstein 1999). The process is backwards looking and is conditioned on the sampled tips. For each lineage the question 'what is the probability that another lineage by chance choose the same parents in lineages above' is asked. This corresponds to 1/n where n is the number of lineages in the parent generation. The population size will affect the tree shape, and the rate of coalescence decreases linearly when the population size n increase (Rodrigo A & Felsenstein 1999), i.e. a larger population has more potential ancestors and hence a lower rate of coalescence. This rate can be affected by the changing transmission rates in infectious diseases, and under this model the population can be assumed to be *constantly* or *exponentially* growing, which will also influence branch lengths in the tree. The coalescent assumes that the sample size *n* is small (Rodrigo A & Felsenstein 1999).

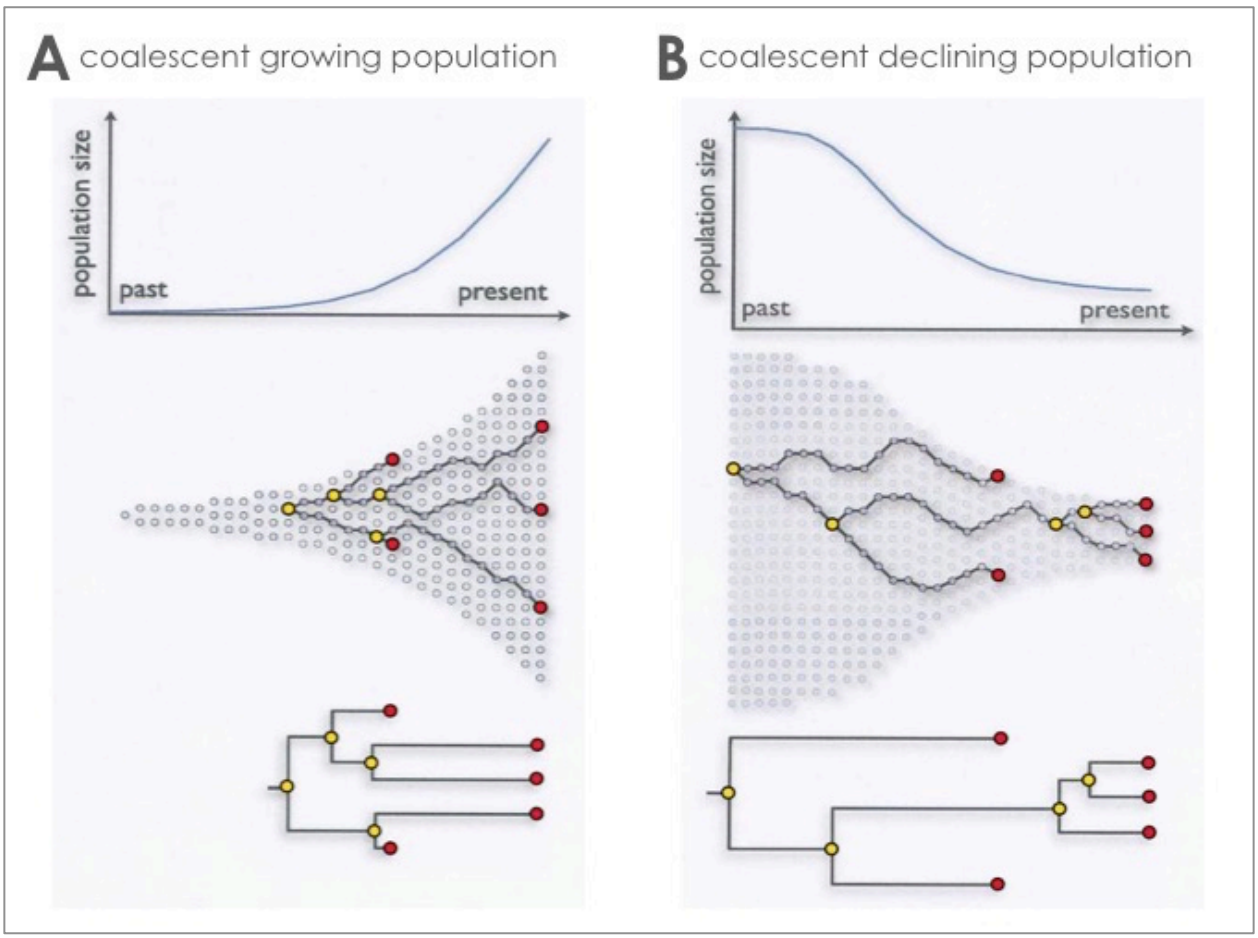


Figure 13. The coalescent theory. The figure shows how the shape of the tree reflects the underlying viral dynamics in a coalescent exponentially growing population (A) and an exponentially declining population (B). When the population size *n* is small (closer to the past/left in tree A, closer to the present/right in tree B) branching events are more common. Picture modified from Andrew Rambeaut, University of Edinburgh (Sainani 2009).

2.4.5.2. Birth-Death models

The birth-death (BD) models describe tree evolution as a branching processes looking forward in time and makes explicit assumptions about the sampling process, e.g. sampling rates and proportions, incorporating these assumptions into the phylogeny. Thus in order to accurately infer epidemiological parameters the sampling proportions must be correctly specified (Li et al. 2014). Like the coalescent the BD models also assume incomplete sampling, but differ in that they condition explicitly on the sampling process, thus being better suited when there is prior knowledge about the sampling. The BD models are especially useful for modelling early phase epidemics with small sample-sizes and large sampling proportions (du Plessis & Stadler 2015).

2.4.5.3. Bayesian skyline models

The skyline extensions of the Bayesian phylogenetic models aim to estimate population dynamics through time, based on a set of molecular sequences collected through time (non-contemporaneous), and to simultaneously incorporate the uncertainties associated with the phylogenetic reconstruction process (Drummond et al. 2005). Briefly, a skyline model assume the population size *N* to be constant in intervals with stepwise changes (Pybus et al. 2000). Plotting the population size *N* to the time, thereby linking these intervals, produces a graph resembling a skyline (fig. 14, panel A). The Bayesian skyline models use MCMC-sampling to create a variant of the generalized skyline plot. Instead of deriving the demographic data from an estimated genealogy, the Bayesian approach takes into account the sequence data and the uncertainties associated with the phylogenetic reconstruction, to generate a posterior distribution of the effective population size through time in addition to credibility intervals for these distributions at any given time (fig. 14, panel B).

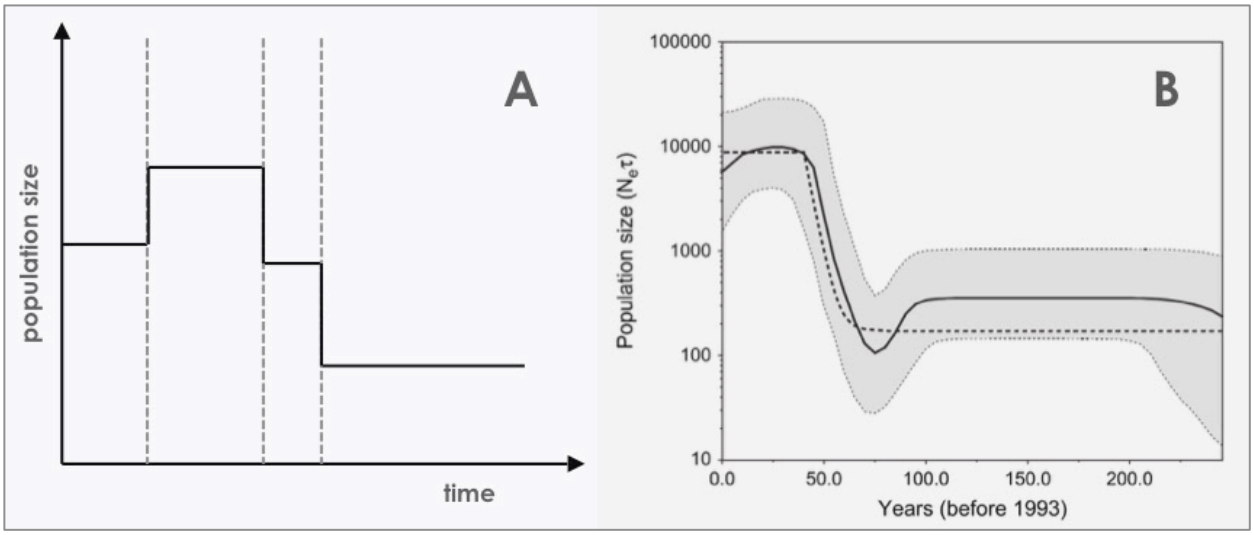


Figure 14. Skyline plots. Panel A: a generalised skyline plot with time on the x-axis and population size on the y-axis. Linking the time-intervals with vertical lines makes the plot resemble a skyline. Panel B: time in years on the X-axis and the N_{et} (product of the effective population size and the generation length in years on the y-axis). The solid line is the mean estimate and the darker gray areas represent the 95% HPD limits. The underlying data in panel B is an alignment of Egyptian HCV sequences, and the graph illustrates the sharp increase in effective number of infections in the early 20th century, which the authors argued to be caused by contamination. Figure on behalf of Drummond et al. (Drummond et al. 2005).

Skyline plots are useful for investigating the development of viral populations through time, and for validating assumptions regarding effects of initiatives, such as control programmes, treatment

regimes, and vaccination programmes. The birth-death skyline (BD-sky) model implemented in BEAST2 (Stadler et al. 2013) can be applied for inferring epidemiological parameters, such as the R_0 when there is reasonable knowledge of the sampling process. However, if this knowledge is lacking the coalescent Bayesian Skyline model is better suited (Drummond et al. 2005).

Chapter 3

MATERIALS AND METHODS

3. MATERIALS AND METHODS

In this chapter, the materials and methods applied in this study will be described in more detail.

3.1. SAMPLE MATERIAL

Each year Kopenhagen Fur receives blood samples (since 2016 on dry blood spot cards) from all mink farms in Denmark, and blood is therefore an easily accessible sample source for genomic surveillance. Blood samples also have the added advantage of being taken without euthanizing the animal as well as containing less interfering host DNA compared to tissue samples. The disadvantage is that the animals need to be viraemic at the time of sampling and hence the number of positives could potentially be low, especially in chronic cases. Taken possible future implementation in mind, blood and serum samples were the main DNA source for the initial PCR development, however due to previous success of others using spleen samples (Jensen et al. 2011) and the fact that most archive samples were in tissue format, these were included in the study as well.

In manuscript 1 it was important to develop a method to capture AMDV strains with different genotypes, and therefore two viral strains with known different phenotypes were chosen as prototypes; the non-virulent cell-culture adapted strain AMDV-G (cell culture isolate, passage 10) obtained from Antigen Laboratory (Glostrup, DK), and the highly virulent AMDV-Utah isolate (also in the form of viral antigen) provided by emeritus Professor Bent Aasted (Copenhagen, DK). In manuscript 2 on the other hand, the goal was to illustrate the benefits of using WGS compared to partial NS1 gene sequencing. Furthermore, the aim was to investigate disease dynamics and the spread in the endemic Northern Jutland region, thus with a geographical focus on Denmark locally to evaluate if WGS added the extra information needed to improve the outbreak investigations. Therefore, the samples for this study were collected from three case farms and from simultaneously infected farms in close geographical proximity to the case farms as well as from farms situated further away in Denmark. AMDV field strains from the archive and the related Gray Fox Amdoparovirus (GFAV) retrieved from Genbank (accession no: JN202450) (Li et al. 2011) were included for rooting.

The archived sample materials (blood, serum and spleen) were kindly provided by Kopenhagen Fur, and represents diagnostic material from the past 10-15 years of AMDV incidences. Additionally, fresh tissue samples were supplied from Polish farms as a part of another related study (Ryt-Hansen, Hagberg, et al. 2017), from a collaboration with Haiko Koenen (DVM, DAC ZuidOost, NL), as well as spleens from wild-mink from Bornholm, kindly supplied by Anne Sofie Vedsted Hammer (Associate

Professor, University of Copenhagen, DK). For further details refer to the sample overview in appendix I.

3.2. DNA-EXTRACTIONS

For the extraction of total DNA from the samples, a manual ethanol-precipitation method and two commercial kits were evaluated: the Qiagen DNAeasy Blood extraction kit and the QIAmp® MinElute Virus Spin Kit (Qiagen, Hilden, DE).

Spleen samples were thawed and 0.01-0.02g was excised from the centre of the specimen using sterile scalpels, followed by DNA-extraction as described below. Blood and serum samples (200µL of each), or 20µL of viral antigen diluted in 180µL PBS, were incubated with proteinase K (Qiagen, Hildren, DE) prior to DNA extraction. The manual ethanol-precipitation was performed as described previously (Gilbert et al. n.d.), where briefly, the DNA was separated into an aqueous phase, from which it was spun down and the pellet was re-suspended in low TE-buffer. DNA extractions using the commercial kits were performed according to the manufacturer's instructions with the following changes: the blood/serum samples were lysed using proteinase K, as it in contrast to the Qiagen protease (a standard serine protease) included in the kit, is not inhibited by EDTA or other blood stabilisers which might be present in blood samples.

The final DNA-elution when using a commercial kit was performed in low TE-buffer, either in 100, 50, or 30µL, to find the optimal concentration for the downstream PCR. For the manual extraction procedure, the DNA-pellet was re-suspended in 2x40µL low TE-buffer followed by incubation on a heating-block at 37°C to reduce the volume to 50µL.

3.3. POLYMERASE CHAIN REACTIONS

The viral DNA was selectively PCR amplified prior to sequencing because of the large amounts of host DNA in the tissue samples, presumed low viral concentrations, the lack of efficient sequencing adaptors for ssDNA, and practical considerations in regards to routine implementation of the workflow in a diagnostic laboratory.

3.3.1. Confirmatory endpoint PCR

Successful DNA-extraction was verified by screening all samples, including a positive (viral antigen) and negative (water) extraction control, with the partial NS1 gene endpoint PCR-primers designed by Jensen et al. (Jensen et al. 2011). The PCR reactions were performed mainly as described before, but with the modification that the buffer and DNA-polymerase was changed to the BioRad SsoAdvanced™ Universal SYBR® Green Supermix (2X) (cat.no. 1725270, Bio-Rad Laboratories Inc., Hercules, CA). The PCR-products were visualised by gel electrophoresis as described previously (Jensen et al. 2011). Successful DNA-extraction was defined as a single PCR-bond of approximately 370 bp for test samples and the positive AMDV control, and a negative lane for the water sample.

3.3.2. Long-range PCR

When this project was initiated previous studies had indicated the presence of secondary structures in the AMDV DNA. Others have chosen strategies such as either amplifying the genome in several overlapping fragments (Li et al. 2012), focussed on shorter fragments (Jensen et al. 2011; Sang et al. 2012), or amplified the genome in several overlapping fragments (Li et al. 2012), which is labour intensive and difficult to automatize. Thus, a long-range PCR covering about 91% of the AMDV-G genome (nucleotide position 98 to 4467 in NC001662) was developed. The use of specific PCR amplification was important for future field applications to avoid abundant host DNA and to attain sufficient amounts of double stranded DNA for preparation of the sequencing libraries. The AMDV-genome was initially amplified in a single fragment, or two or three overlapping fragments. Due to the time constraints and due to the 3'-part of the genome sometimes being more challenging to amplify, it was decided in the difficult cases only to amplify the first 2/3 of the genome, i.e. using primers F1 and R2 to generate what is referred to as "fragment A" (fragm. size 3186, appendix I). The final PCR primer-sequences listed in table 2 were designed using the Primer3 software (Ye et al. 2012). AMDV-G (accession no. NC001662) (Bloom, Alexandersen, et al. 1988) was used as reference genome for primer F1-R3, while field strains were used for designing R5.

Long-range PCR

primer	positions (NC001662)	primer sequence (5'-3')	amplicon size (bp)
F1	77-97	CGCTTCGCGCTTGCTAACTTC	
R1	1814-1793	GCTCTGCGTGAGCGTTTGTTTC	
F1+R1	77-1814		1735
F2	1502-1525	CCGGGGGGCACTGGAAAAACCTTG	
R2	3317-3296	GCAGAGAGGAGGTAGCCCCAAG	
F2+R2	1502-2217		1816
F3	2934-2953	GCGTCGTTACAGGTTGCTTT	
R3	4467-4448	TTAATCCGCCACTTTCTGGT	
F3+R3			1534
R5	4462-4439	CCGCCACTTTCTGGTAAAATAAGG	
F1+R2			3240
F2+R3			1946
F1 + R3			4390
F1 + R5			4385

Real-time PCR

primer	positions (NC001662)	primer sequence (5'-3')	size (bp)
A3F	2836-3856	ACTTAACTGCGTCGTTACAGG	21
A2R	2908-2927	AACAAAGCCCAGTGTTTCCC	20
A2P-f	2874-2900	CCGGGGGGCACTGGAAAAACCTTG	24
amplicon	2836-2927		91
oligo	2826-2937		121

Table 2. PCR primers designed during this project. Forward and reverse primers are indicated F and R, respectively. Expected amplicon sizes in base-pairs are indicated in addition to primer positions relative to the NC001662 AMDV-G reference genome.

Long range PCR reactions were setup as described in Manuscript 1: 25µL GoTag® Long PCR Master Mix (cat.no. M4021, Promega, Madison, WI), 2µM primer F and R, 5µL DNA template, and distilled water up to a total sample volume of 50µL. An annealing-temperature gradient was run and the temperature that generated the clearest PCR-bonds of correct size with the least amount of primer-dimers was selected as optimal. Final cycling conditions were; initial denaturation at 95°C for 2 min, 38 cycles of 30s denaturation at 95 °C, 20s annealing at 58 °C, and 30s extension at 72 °C, followed by a

final extension step for 10 min at 72 °C. All PCR reactions were performed in a Bio-Rad CFX96 Touch instrument (Bio-Rad Laboratories, Inc., Hercules, CA). A positive (AMDV-G DNA) and a negative (ddH_2O) PCR-control were included in all runs to verify correct mixing of the master mix and no cross-contamination, respectively. Ten µL of each PCR product was analysed on a 1% agarose gel stained with ethidiumbromide or SYBR-safe together with a 1kB plus DNA ladder (Invitrogen, Carlsbad, CA) to confirm successful amplification of bands with the expected size.

3.3.3. Real-time PCR

For real-time PCR it is important to consider that shorter DNA fragments are easier to amplify and hence will improve efficiency. Factors that adversely affect PCR-efficiency are e.g. the presence of secondary structures in the target region, inappropriate GC-content (preferably around 50%) or simply poorly designed primers, and if applicable probes. It was essential for the real-time assay to detect all known AMDV strains equally well, and therefore four different genomic regions were targeted. For each region three different slightly overlapping primer-pairs were designed using the Primer3 software (Ye et al. 2012) implemented in Geneious v.7.1.5 (Kearse et al. 2012) with a whole genome sequence alignment with representatives from the three known AMDV genotypes included (appendix I). The real-time PCR primer- and probe sequences are listed in table 2. In each of these four regions a primer-matrix with all possible combinations of forward and reverse primers was ran using SYBR®-green chemistry, and the two best performing primer-pairs, assessed by gel electrophoresis, were selected for final bench-marking together with their corresponding probes.

3.3.3.1. Double stranded DNA intercalating chemistries

Double stranded DNA intercalating chemistries such as SYBR® and EvaGreen contains a dye that upon incorporation into double-stranded DNA emits light. The signal accumulation during each PCR cycle is monitored at a predefined wavelength (here around 500nm) and the cycle number where the amount of signal exceeds the background-threshold is referred to as the quantification cycle (C_q) or C_q-value. Since the incorporation of dye is unspecific to any dsDNA the analysis relies on primer design for specificity. An additional final step was added to the cycling protocol, after the last round of amplification, i.e. a melting-curve step (fig. 7). The decrease in fluorescence signal was monitored during small temperature increments to create a signal peak followed by a decrease at the temperature where the main bulk of PCR-product was denatured (i.e. the target DNA). This act as a specificity control ensuring the majority species in the reaction really is one species, as multiple peaks would be observed if two or more species existed.

3.3.3.2. Hydrolysis probe chemistries

Hydrolysis or dual labelled probe chemistries, sometimes referred to as TaqMan®-assays, have in addition to assay specific primers, a specific probe designed to bind a certain site on the target DNA (between the primers). The probe was in the 3'-end labelled with a 6-carboxyfluorescein (6-FAM) fluorophore that emits light at a specific wavelength (517nm), and in the 5'-end with a black hole quencher (BHQ-2) that hinders the fluorophore to fluoresce as long as both are bound to the intact probe. The probe is designed to anneal (bind) to the target DNA at a temperature approximately 10 °C higher than the primers, hence binding before the primers during cycling. When the DNA polymerase elongates the DNA strand starting at the primer binding sites it will reach and cleave the probe due to its 5' to 3' exonuclease activity. This results in separation of the fluorophore from the quencher and allows for the emission of light. The PCR machine registers this increase in signal.

3.3.3.3 Assay performance

Real-time PCR performance was measured by looking at the assays efficiency and reproducibility (Bustin et al. 2009). PCR efficiency is a measure for how well the target is amplified, and ideally each target is copied once for every PCR-cycle thereby exponentially doubling the amount of target in the reaction, corresponding to an efficiency of 100%. By plotting the C_t-values of a 10-fold dilution of a template with known concentration to its increase per cycle the slope should be close to -3.322. A smaller slope (e.g. -3.8) gives a <100% efficiency and is a sign of experimental limitations such as inhibitors in the sample. A larger slope (e.g. -3.2) equals an >100% PCR-efficiency, which can occur when values are measured in the nonlinear phase of the reaction, due to inhibitors or an inaccurate dilution series.

3.4. DNA SEQUENCING

3.4.1. Sample preparation

The PCR products were purified according to the manufacturer's instructions using the QIAquick® PCR Purification Kit (cat.no. 28104, Qiagen Hildren, DE) or the QIAquick® Gel extraction kit (cat.no. 28704, Qiagen Hildren, DE) depending on if there was a single product or multiple bands on the gel. Briefly, when a single band was present on the gel 40µL of PCR-product was added to 250µL buffer and if needed, the pH was adjusted with 3M sodium acetate before loading samples to the spin columns and proceeding with the extraction according to the manufacture's protocol. If multiple bands were present on the gel 2x20µL of each PCR product was loaded into adjacent wells on a 0.8%

agarose gel stained with SYBR®-safe (cat.no. S33102, Invitrogen, Carlsbad, CA), quickly visualized under UV-light and fragments of correct size were manually excised from the gel with a scalpel. The gel slices were weighted and dissolved in three volumes of buffer and subsequently extracted according to the manufacturer's protocol. Regardless of the purification method the PCR-products were finally eluted in low TE buffer before sequencing.

3.4.2. Next generation sequencing

The purified PCR-products were directly submitted for NGS at DTU (Technical University of Denmark) Multi-Assay Core (Lyngby, Denmark) who prepared the 400bp sequencing libraries and sequenced the samples on a 318-chip using the Ion Torrent PGM™ (Life Technologies, Carlsbad, CA) according to the manufacturer's instructions. Briefly, sequencing is based on registration of a current created from the release of protons upon incorporation of nucleotides into a growing DNA strand. The Ion Torrent technology was selected because it was easily accessible at the DTU Core facility (DMAC) in Lyngby, and it is cheap, fast, and provides reasonable read-lengths (usually 150-400bp), all-important factors for a future intended routine implementation and validation.

3.4.3. Sanger sequencing

The genomic region between positions 2,470 to 2,520 displayed low NGS read coverage. Since the Ion Torrent is known to struggle reading homopolymeric and GC-rich regions (Quail et al. 2012), PCR-products spanning this region were Sanger-sequenced using primers F2 and R2 (table 2) to verify the Ion Torrents findings. The following samples were tested in this manner: AMDV-G, AMDV-Utah, and the field strain no. 316 (appendix I). Each PCR-product was divided into two 10µL aliquots of PCR-product and added 5µM of either forward or reverse primer and Sanger sequenced by LGC Genomics (Berlin, DE).

3.5. SEQUENCE ANALYSIS

3.5.1. Raw-data structure

The output from DNA sequencing is a raw-data file containing the base letters and their associated quality scores. For the Ion Torrent sequencer the raw-data is provided in a text-based format, a so-called fastq-file, which can, given the massive parallel sequencing, be very large and puts requirements on the computing capacity both for transfer, storage, and analysis. A fastq-file (fig. 16) contains four rows per read; on the first row is an identifier beginning with an '@' followed by a referral to the read itself (e.g. seq_ID), the second row contains the raw sequence letters

(nucleotides), the third row begins with a '+' and can optionally be followed by the same sequence identifier, and the fourth row an ASCII-encoded quality score for each base. The quality score is a logarithmically linked to the probability of an incorrect base call ($Q=10\log_{10}P$), where e.g. a quality score (Q) of 20 corresponds to 1 of 100 bases being incorrect (i.e. 99% accuracy).

Figure 16. Fastq- and fasta-formats illustrated. Fastq-file with four lines: identifier, raw bases, '+', and ASCII-encoded quality scores. Fasta-file with header starting with '>' followed by the bases on a new row.

Most data processing was performed in the command line in a Unix environment using the Danish National Supercomputer for Life Science Computerome (https://computerome.dtu.dk). An overview of the pipeline can be found in figure 17.

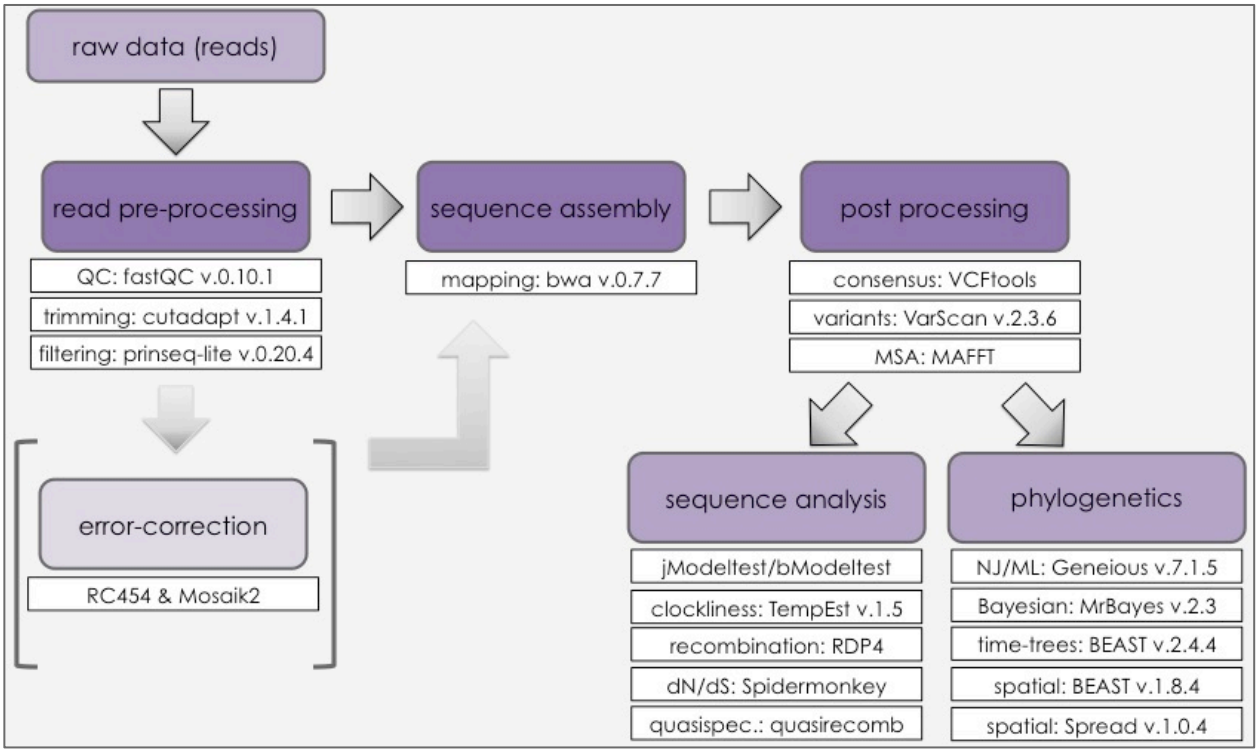


Figure 17. NGS-pipeline flow-chart. Overview of the analysis steps: the raw data (reads) goes into pre-processing (quality check (QC), filtering and trimming), sequence assembly (either by mapping to reference or by *de novo* assembly), and selected post-processing alternatives (variant calling, generating a consensus) and phylogenetics. In some instances, an error-correction step preceded the assembly step.

3.5.2. Data pre-processing

The overall aim when pre-processing the data is to control and correct the quality of the reads and to remove noise in order to achieve the most accurate results in the downstream analyses. Primer sequences were removed using Cutadapt v.1.4.1 (Martin 2011), and reads were trimmed according to quality score (>20) and read-length (100-400) using Prinseq-lite v.0.20.4 (Schmieder & Edwards 2011). The input files were raw data from NGS in fastq-format, and the output was renamed with jorunal-id, and *_one (if amplified in one fragment).

```
# cutadapt -O 19 -e 0.05 -g primerF -g primerR -g primerF2 -g primerR2 rawfile.fastq -m 1 | prinseq-lite.pl -fastq stdin -out_good stdout -out_bad null -trim_to_len 400 -trim_left 10 -trim_qual_left 20 -trim_qual_right 20 -min_len 100 -min_qual_mean 20 > *.cut-trim.fastq
```

Summary quality statistics were generated prior to and after trimming using FastQC v.0.10.1 (Andrews 2010).

```
# fastqc -o fastqc_fl_12 *rawfile.fastq
# fastqc -o fastqc_fl_12 *cut-trim.fastq
```

Raw-data generated by Sanger sequencing was imported into Geneious 7.1.5. (Kearse et al. 2012) where low quality bases in the ends and primer-sequences were removed manually. The forward and reverse strains were merged into a consensus to be used for further analysis.

3.5.3. Error correction

The Ion Torrent PGM used in this project registers the release of protons upon incorporation of bases during sequencing. The signal is proportional to the number of incorporated bases and this creates known problems accurately reading homopolymeric regions and carry forward/incomplete extension. ReadCorrect454 (RC454) (Henn et al. 2012) is a software package developed to correct reads for these known 454 problems. The Ion Torrent PGM sequences and generates data in a similar manner to the 454 and others have applied RC454 on Ion Torrent reads with good results (Fahnø 2014). RC454 runs together with the Mosaik2 aligner (Lee et al. 2014) which takes the reads and aligns them to a given reference (here NC001662 cut between position 98-4466). Mosaik2 uses the neighbourhood quality standard (NQS), that is, for each base it not only uses its own Phred score but also takes into account the scores of the adjacent bases (Lander et al. 2000). The default values were used, thus a quality score of q=20 and n=5 neighbours around a given base must pass a quality score of q'=15. ReadCorrect454 (RC454) (Henn et al. 2012) takes the reads (split into reads and quality files), a consensus assembly for those reads and their alignment to the qlx-format assembly generated by the wrapper script runMosaik2.pl, and corrects them for difficulties with homopolymeric regions. It corrects for any indels that breaks a reading frame, unless it occurs in more than 25% of the reads. RC454 uses Mosaik2 to align the corrected reads between each step.

Firstly, the fastq-file (containing the reads of interest) is split into a fasta and a qual file (using the "o"-flag will automatically append "*.fasta" and "*.qual" to the output):

```
# python convertFastq2FastaQual.py --i *.fastq --o *.cut-trim
```

The reads are aligned to a chosen reference (or assembly) using Mosaik2. Use the "qlx"-flag to generate qlx-format and the "param454"-flag to customize the algorithm for 454 data (e.g. gap penalties):

```
# perl runMosaik2.pl -fa *.cut-trim.fasta -qual *.cut-trim.qual -ref reference.fasta -o *.cut-trim-RC454 -qlx -param454
```

RC454 takes the aligned reads from the step above (qlx), the original reads, and the reference/assembly and outputs with "bam"-flag to return .sam and .bam formats.

```
# rc454.pl *.cut-trim-RC454.qlx *.cut-trim.fasta *.cut-trim.qual ref.fasta *cut-trim-RC454 -bam
```

The cleaned reads (in fasta and qual formats) are gathered to a final fastq-file containing the corrected reads (automatically makes fastq):

```
# python convertFastaQualtoFastq.py --fasta *.cut-trim-RC454_cleaned.fasta --qual *.cut-trim-RC454_cleaned.qual
```

The final fastq file was renamed (optional):

```
# cat *.cut-trim-RC454_cleaned.fastq > *.RC454.cut-trim.fastq
```

The error-correction step was applied initially, however, manual inspection of raw-reads and nucleotide comparisons between the same sequences processed with and without this step showed that it did not correct the reading-frames and it was therefore removed from the pipeline.

3.5.4. Sequence assembly

The pre-processed reads were, regardless if error-corrected or not, assembled to a consensus sequence by mapping using the software Burrows Wheeler Align (BWA) v0.7.7 (Li 2013) and the AMDV-G reference genome (NC001662) for guidance. BWA is designed to map low-divergent sequences to a reference genome and throughout this study the BWA-MEM option was applied, as it is fast, accurate, and suitable for longer reads (Li 2013).

```
# bwa mem NC_001662_98_4466.fasta *.cut-trim.fastq | samtools view -Sb -> *.RC454.cut-trim.bam
```

The bam-files were sorted and indexed using Samtools with the flagstat option (Li et al. 2009).

```
# samtools flagstat *.RC454.cut-trim.bam
# samtools sort *.RC454.cut-trim.bam *.RC454.cut-trim.sort
# samtools index *.RC454.cut-trim.sort.bam
```

The read-depth for each nucleotide position along the genome was computed using bedtools genomecov (Quinlan & Hall 2010), and outputted in CSV-format. The coverage for each position in the genome was computed using Microsoft® Excel® for Mac 2011 version 14.6.9 (www.microsoft.com) and plots in PDF-format was generated in R version 3.2.2 (R core team 2015).

```
# bedtools genomecov -ibam *.RC454.cut-trim.sort.bam -d > *.RC454.cut-trim.sort.bam.cov
# rscript coverage.r
```

3.5.5. Post-processing analyses

The mpilup module in Samtools takes the sorted alignment file (*.sort.bam), and creates a pileup of overlapping reads for each position. This pileup was piped into VarScan v.2.3.6 (Koboldt et al. 2012), which called variants and short indels with a base frequency cut-off > 0.5 for each position. VCFtools v.0.1.12a (Danecek et al. 2011) was used for converting the SNP calls and indels into a consensus-

sequence for further use in downstream analyses.

```
# samtools mpileup -E -f NC_001662_98_4466.fasta *.RC454.cut-trim.sort.bam | java -jar
VarScan.v2.3.6.jar mpileup2cns --min-var-freq 0.5 --output-vcf 1 --p-value 0.01 --min-avg-qual 20 |
bgzip > *.RC454.con.vcf.gz
tabix -p vcf *.RC454.con.vcf.gz
# cat NC_001662_98_4466.fasta | vcf-consensus *.RC454.con.vcf.gz > *.RC454.con.fasta
```

The input for phylogenetic analyses is multiple sequence alignments (MSA). The consensus sequences in fasta-format were combined into a single fasta-file and aligned at nucleotide level using MAFFT (Katoh & Standley 2013) with the auto-setting, which searches various penalising options and selects the most suitable for the data, and were then converted to nexus format with the python-script Seqconverter (Anders G. Pedersen 2012) by running the following commands:

```
# cat *.fasta > dataset.fasta
# mafft --auto dataset.fasta > dataset_aln.fasta
# seqconverter -I fasta -O nexus dataset_aln.fasta > dataset_aln.nex
```

When certain regions of the genome were to be analysed individually, the partition between its start and stop coordinates were cut out using subseq-flag in the python-script Seqconverter (Anders G. Pedersen 2012).

```
# seqconverter --subseq=start,stop dataset_aln.nex > dataset_cut_aln.nex
```

3.5.6. Model-testing

For each alignment (i.e. dataset) the best fitting substitution model was selected using jModeltest (Darriba et al. 2012) or the package bModelTest (Bouckaert 2015) implemented in BEAST2. When using jModeltest, the best fitting model was selected based on the Akaike Information Criterion (AICc), which is the AIC corrected for a finite sample size, and therefore penalises harder on large numbers of parameters than AIC. The AIC measures the relative quality of a model compared to other models for a given data set, and it provides an estimate of the information that is lost when a given model is used to represent the process generating the data. It tells nothing about the quality of the model or if it fits poorly. The AICc is recommended when the sample size n is small and/or when the number of free parameters k is large. AICc converges to AIC when n is large and therefore it can be advised to always use AICc (Burnham & Anderson 2002). The bModelTest uses a Bayesian approach with MCMC sampling to explore different states representing different substitution models, while simultaneously estimating the other model parameters. It has however, been shown that phylogenetic analyses often are relatively robust to the choice of nucleotide model (Drummond & Bouckaert 2015), which was verified by running different models to generate similar values.

The aim with model selection is to determine if the more parameterised model fits the data (alignment) significantly better taken into consideration the added number of parameters. The usefulness of these tests has been debated (Bouckaert 2015) as e.g. differences in selective pressure over the genome might result in different nucleotide substitution models (Melorose et al. 2009). Therefore, a more empiric step-wise approach can be beneficial (personal communication), hereunder investigating for potential recombination before making the phylogenetic analysis, run the analysis, and perform "sanity check" on the output (i.e. log-files and trees).

3.5.7. Investigating clocklikeness

Because the BEAST software assumes a molecular clock rate larger than zero it is advised to investigate if the dataset exhibits clocklikeness behaviour before applying a molecular clock model. TempEst (Rambaut et al. 2016) plots the genetic distance of an ML-tree to the sampling-dates. A linear relationship with small residuals indicates that evolution has occurred with a strict clock-like behaviour, while larger residuals suggest a relaxed molecular clock. Non-linear trends on the other hand, suggests evolutionary rates have changed through time, while no trend at all, implies there is no or little temporal signal and that data is unsuitable for molecular clock models. Large y-axis residual indicates there is a problem with the sequence itself, such as low quality, bad assembly, or an alignment error. It could also be error in the phylogenetic inference, or a biological process such as recombination. Large x-axis residuals indicate the specified sampling-date does not match the observed genetic distance and could be caused by e.g. mislabelling, biological contamination or error in phylogenetic inference (Rambaut et al. 2016).

3.6. PHYLOGNETIC ANALYSES

Phylogenetic analyses aims to reconstruct the history of given species (or set of samples) and put this on an evolutionary scale (e.g. years or substitutions per site per year). The BEAST-packages (described below) were used for all phylogenetic inferences where sampling-dates and other metadata were included, while quicker and exploratory tests of phylogenetic models and relationships were mainly performed in MrBayes v.3.2.3 (Ronquist et al. 2012) as described below.

3.6.1. Maximum likelihood phylogenies

Maximum likelihood (ML) trees were constructed using the ML-plugin in Geneious v.7.1.5 (Kearse et al. 2012). Briefly, the goal with ML is to maximise the probability of observing the sequences in the dataset, given the model. This is done through stepwise determination of the models parameter-

values, such that the likelihood of the observed sequences increases. The most optimal tree can be found by changing the structure of a given tree according to fixed rules, thereby generating a new set of trees (the neighbourhood) and compute the likelihood for each of these trees. The tree with the highest likelihood is then used to repeat the procedure until no better tree is found. The reliability of the ML trees, i.e. the self-consistency of the data, was evaluated by bootstrapping the alignment *N* times (sampled with replacement) and for each alignment a new ML-tree was calculated. Bootstrap values are associated with branches, not the nodes, and commonly used cut-off values are in the range 0.5-0.7 meaning that a branch should appear at the same place in more than 50-70% of the trees in order to be considered valid (Melorose et al. 2009).

3.6.2. Estimating phylogenetic relationships using MrBayes

Phylogenetic relationships were inferred in a Bayesian framework with Markov-chain Monte Carlo (MCMC) sampling using MrBayes version 3.2.3 (Ronquist et al. 2012). Prior to performing a Bayesian phylogenetic analysis, a model needs to be described for the data, and the uncertainties about the parameters are specified with probability distributions (the priors). These prior distributions are then updated, based on MCMC sampling of the data, to posterior distributions. The DNA-substitution models were specified according to the model test results and when applicable, the proportion of invariable sites was estimated from the data with four gamma distribution rate categories. The prior distributions were kept to the default settings. Each MCMC consisted of four parallel chains (default) and each analysis was initially run for 10M generations for quick visualisations, followed by two or more longer runs (often 50M) to ensure convergence not was due to local optima. MrBayes (and BEAST) gathers the MCMC samples in a tab delimited plain text file with one row per sample (log-file). These log-files were accumulated into frequency distributions providing estimates of the marginal posterior probability distribution for each parameter using the designated tool Tracer v.1.6.0, distributed with the BEAST-packages (Drummond et al. 2012; Bouckaert et al. 2014). The first 25% of the samples were discarded as burn in, effective sample size (ESS) values above 200 for all parameters, and standard deviation of split frequencies below 0.001 were considered as indications that the MCMCs had converged successfully. Summary trees were generated using a 25% burn-in.

Command-line example specifying a model with an HKY-model of substitution, a proportion of invariable sites, a gamma-rate distribution, sampled by an MCMC of 10M generations:

```
# mb
# exec dataset_aln.nex
# lset=2 ncat=Invgamma
# mcmc ngen=10000000 Samplefreq=1000 savebrlens=yes
# sumt relburnin=yes burninfrac=0.25
# sump relburnin=yes burninfrac=0.25
# quit
```

3.6.3. Divergence and time-calibration and estimating the molecular clock

BEAST is acronym for "Bayesian evolutionary analysis by sampling trees" and is a software package providing a phylogenetic framework for simultaneously estimating phylogenies and their parameters. BEAST currently exists in two different versions supported by somewhat different developers: BEAST v.1 (Drummond et al. 2012) and its complete rewrite BEAST2 (Bouckaert et al. 2014). In essence, BEAST allows the parameters of interest, e.g. mutation rate and population size, jointly to be estimated from temporally spaced sequences by simultaneously incorporating their genealogical uncertainty. This is done in a Bayesian framework and the sampling is performed using Markov Chain Monte Carlo (MCMC) integration. In order to run, BEAST needs an XML-file containing the data (the alignment), metadata (if applicable), substitution, clock- and tree model, prior settings, and a user-determined number of MCMC iterations. The XML-file can be set up using the graphical user interface "Beauti" distributed together with the BEAST-package or by manual text editing. BEAST generates two output files of interest: a log file and a tree file (described in more detail below).

Time-calibrated phylogenies showing the relationships between the sequences in a tree with branch lengths on a calendar timescale were constructed using BEAST v.2.4.4 (Bouckaert et al. 2014). The tree topology, the age of the internal nodes, the rate of evolution, and the fit of the substitution model, were simultaneously inferred by fixing the tips (sequences) to their sampling-dates and then applying MCMC sampling to determine the ages of the internal nodes of the trees. The length of the branches were thus displayed in units of time, and mapped to an expected number of substitutions per site according to a vector of molecular evolutionary rates (the clock).

The alignments used for the phylogenetic analyses in MrBayes described above, contained outgroup sequences (AMDV-G or GFAV), to facilitate rooting of the trees, defining internal nodes, and identifying major clusters. As a rule of thumb, the outgroup should have sufficient evolutionary distance to root the tree, but not be so different that long branch-lengths are introduced, since this may affect estimation tasks, especially in Bayesian time-line analyses (Drummond & Bouckaert 2015). Thus, the evolutionary distant outgroup sequences (AMDV-G and GFAV), in addition to sequences lacking sampling-dates, were removed from the time-calibrated analyses.

Both strict and relaxed molecular clocks, in addition to coalescent constant and coalescent exponential tree population priors were tested, and all runs were sampled with MCMC's run for 10-100M generations to obtain estimates of the posterior distributions. Each MCMC was run at least twice to verify convergence to similar values. The coalescent was applied as tree prior, as it is suitable when there is no or little prior knowledge about the sampling process (i.e. sampling proportions and rates). Furthermore, the coalescent assumes incomplete sampling, which makes it especially useful for inferring epidemic history from pathogen sequences for diseases with mild or asymptomatic infections, and for which case-based surveillance data underestimate prevalence (Li et al. 2014). A gamma distribution was applied to describe the collected substitution rate for all sites in the alignment (Yang et al. 1996). A gamma distribution with a small alpha shape parameter (e.g. 0.1), results in a curve resembling an exponential, meaning that most sites are assumed to have a low mutation rate and the tapering tail tells that a few sites have a higher rate. A gamma distribution with a large alpha (e.g. 10) would peak around 10 and almost resemble a normal distribution, and therefore assume a close to normal distribution of substitution rates across sites. That is, most sites mutate with a rate around 10, and some sites slower, and other faster. The L-shaped prior was applied whenever a gamma or invariable gamma model was indicated by the model test.

3.6.4. Estimating viral population dynamics through time

The skyline extensions implemented in BEAST2 was used to estimate past population dynamics through time based on samples of molecular sequences (Drummond et al. 2005). There are two main approaches for calculating the tree-prior probability: the coalescent or the birth-death model assumption. As a rule of thumb, birth-death models are better for modelling early phase epidemics with small sample-sizes and large sampling proportions, while coalescent approaches better account for complex population dynamics (du Plessis & Stadler 2015). The BD approach on the other hand, can be useful for smaller and densely sampled outbreaks, but care should be taken as in order to accurately infer epidemiological parameters, the specified sampling proportions must be correctly assigned (Li et al. 2014). Referring to the section above, the coalescent version of the Bayesian Skyline model was applied in the present study, as it is less dependent on detailed prior knowledge about the sampling process.

3.6.5. Tree visualisations

In addition to log-files, a phylogenetic analysis outputs tree-files, which contain the posterior sample of trees. For each run these were summarised into a maximum clade credibility (MCC) tree using TreeAnnotator v.2.4.2 (Bouckaert et al. 2014). When necessary due to large file-sizes LogCombiner

v.2.4.2 (Bouckaert et al. 2014) was used for summarizing the tree log files, and FigTree v.1.4.2 (also distributed with the BEAST-package), were used for tree manipulations such as rooting and visualisation. To facilitate visualisation, the root sequence was sometimes removed from the MCC trees using the custom python-scripts treerooter.py and treecutter.py (A.G. Pedersen 2012).

Uncertainties in tree topology were furthermore visualised using DensiTree (Bouckaert 2010), distributed with the BEAST-package. This software allows for qualitative analysis of a Bayesian phylogenetic trees by creating a visual overview highlighting well supported clades, distributions of MRCA's, and topological uncertainties (Bouckaert 2010). Areas where a lot of trees agree will be depicted as more dense and clear, while unsupported areas will appear diffuse. Furthermore, sanity checks were performed regularly, i.e. 'looking at the tree and applying prior knowledge, if existing'.

3.7. GENOMIC VARIATON

3.7.1. Sequence diversity

Intra-farm diversity was estimated by calculating the mean pairwise distance and its corresponding standard error (SE) between the sequences collected from each farm using a custom Python script (Anders Gorm Pedersen 2012).

3.7.2. Recombination

The presence of recombination between sequences in a dataset can impact the phylogenetic analyses. E.g. a recombinant sequence A could be most similar to its ancestor B in the 5'-end and most similar to ancestor C in the 3'-end (fig. 18). Thus, a phylogenetic tree created from such alignment would differ depending on which part of the sequence/alignment was used for the inference.

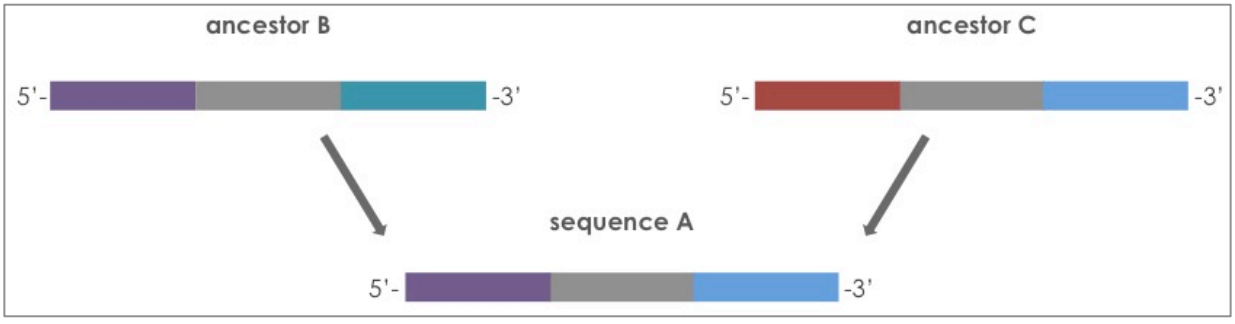


Figure 18. Recombination. The recombinant sequence A is most similar to its ancestor B in the 5'-end, while most similar to its ancestor C in the 3'-end.

The full-length AMDV isolates were investigated for recombination using RDP4 (Martin et al. 2015), a software that simultaneously apply a range of recombination detection methods. The most common approach is the "average over window-based approach". Essentially the alignment is partitioned, a measure for relatedness is calculated for each partition (window), and these measures are compared in order to identify which of the partitions might have recombined.

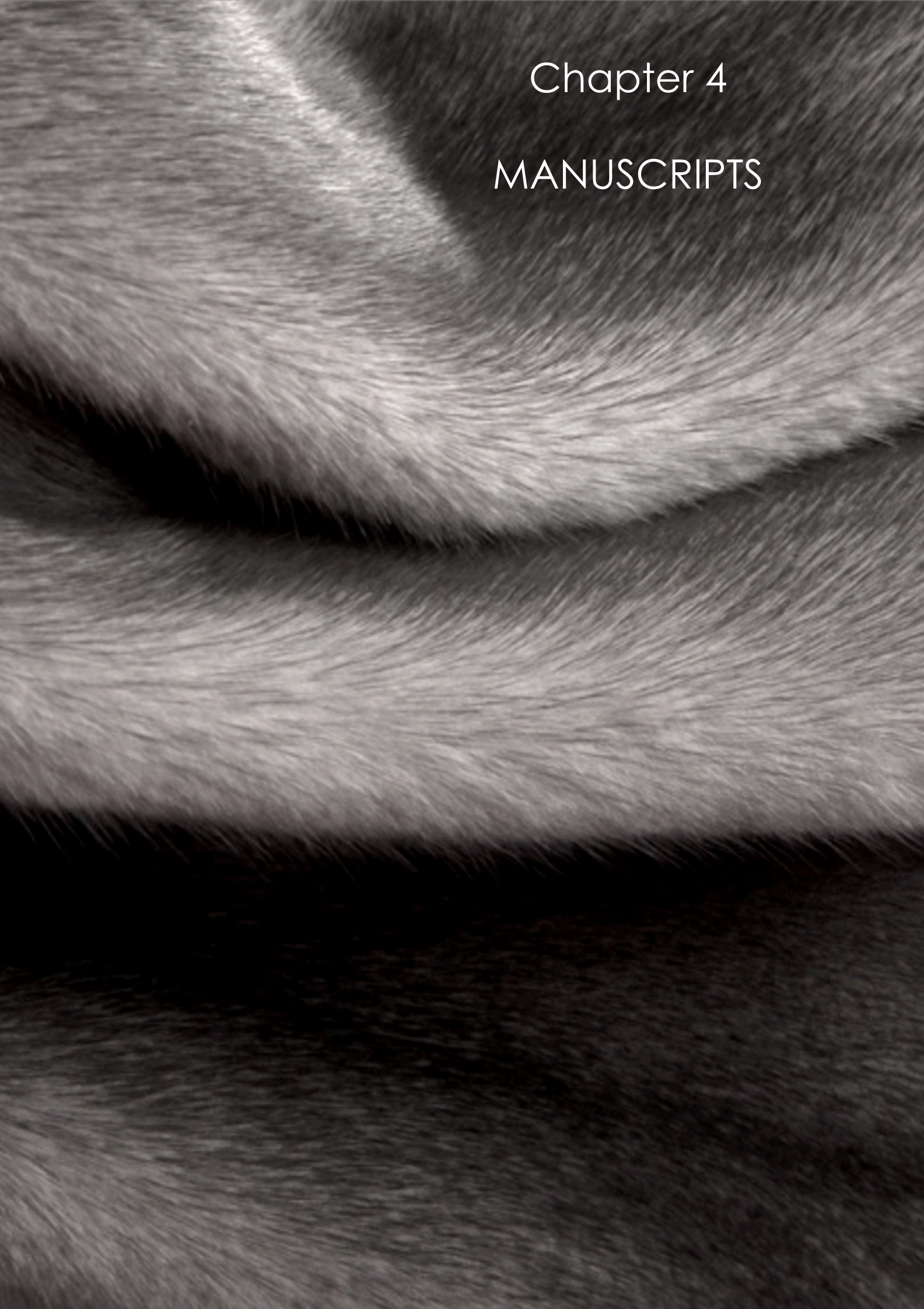
Caveats for applying exploratory recombination methods are e.g. when reference sequences (i.e. known recombinants) are lacking, as this might reduce the statistical power (Melorose et al. 2009), or when the sequence divergence is low (Posada et al. 2002). Most recombination methods compute the pairwise genetic distance between all sequences in the alignment with the assumption that closer related sequences are more similar, which is not always true. The relative distance between non-recombinant sequences should remain constant over the alignment, otherwise it is a sign of recombination and breakpoints can be identified. It can however be difficult to asses which sequence is jumping (Melorose et al. 2009).

3.7.3. Selection pressure

The selective pressure was estimated by calculating the dN/dS ratio, i.e. the number of non-synonymous substitutions per non-silent site (dN) compared to the number of synonymous substitutions per synonymous site (dS). A ratio < 1 is called purifying or negative selection, and indicates that selection has acted against deleterious substitutions with reduced capacity to complete viral replication or other stages in the life cycle. A ratio > 1 on the other hand, indicates that the main driver is positive selection, i.e. that selection or genetic drift has favoured certain amino acids. Despite being a widely used tool, the d_N/d_S ratio is best suited for assessing selection *between* viral populations, as it is somewhat insensitive when highly similar isolates (as in the present study) are compared (Kryazhimskiy & Plotkin 2008).

For each of the two major genes, NS1 and VP2, their *overall ratio* of synonymous and non-synonymous substitutions were calculated using the Single Likelihood Ancestor Counting (SLAC) implemented on the site Datamonkey (www.datamonkey.org). The selective pressures acting at individual codons were assessed separately for codon-alignments of the NS1 and VP2 genes using four designated methods implemented at the web interface Datamonkey (Pond, S. L., Frost 2005), i.e. the Single Likelihood Ancestor Counting (SLAC), the Fixed Effect Likelihood (FEL), the Internal FEL (IFEL), the Random Effects Likelihood (REL). These are all likelihood-based methods that performs somewhat differently: e.g. SLAC is considered less sensitive and FEL is known for overestimating the number of sites, and it

is common practice to accept a codon as being under selection when predicted by at least two of the above mentioned methods (Canuti et al. 2016). For more details refer to Pond, S.L. and Frost (2005).



Chapter 4

MANUSCRIPTS

4. MANUSCRIPTS

This chapter contains the manuscripts produced as a result of this thesis (both published and in preparation).

4.1. MANUSCRIPT 1

A fast and robust method for whole genome sequencing of the Aleutian Mink Disease Virus (AMDV) genome.

Status: Published in Journal of Virological Methods (JVM), April 2016.

(page numbers are relative to paper)

Journal of Virological Methods 234 (2016) 43–51

Contents lists available at ScienceDirect

Journal of Virological Methods

journal homepage: www.elsevier.com/locate/jviromet

A fast and robust method for whole genome sequencing of the Aleutian Mink Disease Virus (AMDV) genome

Emma E. Hagberg[a,b,*], Anders Krarup[a], Ulrik Fahnøe[c,1], Lars E. Larsen[c], Rebekka Dam-Tuxen[a,2], Anders G. Pedersen[b]

[a] Kopenhagen Diagnostics, Kopenhagen Fur, Glostrup, Denmark
[b] Department of Systems biology, Technical University of Denmark, Lyngby, Denmark
[c] National Veterinary Institute, Technical University of Denmark, Frederiksberg, Denmark

Article history:
Received 9 October 2015
Received in revised form 23 March 2016
Accepted 23 March 2016
Available online 6 April 2016

Keywords:
AMDV
PCR
NGS
Whole genome sequencing

A B S T R A C T

Aleutian Mink Disease Virus (AMDV) is a frequently encountered pathogen associated with commercial mink breeding. AMDV infection leads to increased mortality and compromised animal health and welfare. Currently little is known about the molecular evolution of the virus, and the few existing studies have focused on limited regions of the viral genome.

This paper describes a robust, reliable, and fast protocol for amplification of the full AMDV genome using long-range PCR. The method was used to generate next generation sequencing data for the non-virulent cell-culture adapted AMDV-G strain as well as for the virulent AMDV-Utah strain. Comparisons at nucleotide- and amino acid level showed that, in agreement with existing literature, the highest variability between the two virus strains was found in the left open reading frame, which encodes the non-structural (NS1–3) genes. This paper also reports a number of differences that potentially can be linked to virulence and host range.

To the authors' knowledge, this is the first study to apply next generation sequencing on the entire AMDV genome. The results from the study will facilitate the development of new diagnostic tools and can form the basis for more detailed molecular epidemiological analyses of the virus.

1. Introduction

Aleutian Mink Disease (AMD), sometimes referred to as Plasmacytosis, is worldwide the most important disease in the mink farming industry. The disease affects mink of all ages and is caused by Aleutian Mink Disease Virus (AMDV), a single stranded DNA virus belonging to the family *Parvoviridae* (Bloom et al., 1980) genus *Amdoparvovirus* species *Carnivore amdoparvovirus* 1. Viral entry is respiratory, oral, or via the placenta (Broll and Alexandersen, 1996). Infection results in a harmful activation of the immune system leading to hypergammaglobulinaemia and systemic vascular diseases and glomerulonephritis. Animal welfare is reduced and infected animals either die due to organ failure or become persistently infected carriers transmitting the virus within and between herds (Decaro et al., 2012). Like other parvoviruses AMDV replicates only in dividing cells where it utilizes the host cell's transcription machinery. Multiple parvoviruses can infect the same host, and this is believed to contribute to the high recombination rate shown for parvoviruses compared to other DNA viruses (Shackelton et al., 2007). AMDV consists of two large open reading frames (ORF's); the left ORF (nucleotide 116–1975) coding for the non-structural (NS) proteins involved in gene regulation and replication, and the right ORF (nucleotide 2241–4346) coding for the viral capsid proteins (VP), and three smaller central ORF's (Alexandersen et al., 1988; Bloom et al., 1988). In Denmark AMDV is a pathogen that is monitored by a mandatory national control program (Danish Executive Order 1447 of 15/12/2009, 2009). Briefly, the program requires all farms to conduct screening of their animals at regular intervals according to the disease status of the region. Positive farms undergo a more intensive monitoring and are encouraged to depopulate followed by a thorough cleaning and disinfection. Given that parvoviruses are highly contagious and very resistant to environmental factors, managing AMDV imposes large costs on the fur

* Corresponding author.
E-mail address: eha@kopenhagenfur.com (E.E. Hagberg).
[1] Current address: Copenhagen Hepatitis C Program (CO-HEP), Department of Infectious Diseases and Clinical Research Centre, Hvidovre Hospital and Department of International Health, Immunology and Microbiology, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark.
[2] Current address: Novo Nordisk A/S, Smørmosevej 17-19, DK-2880 Bagsværd, Denmark.

http://dx.doi.org/10.1016/j.jviromet.2016.03.010

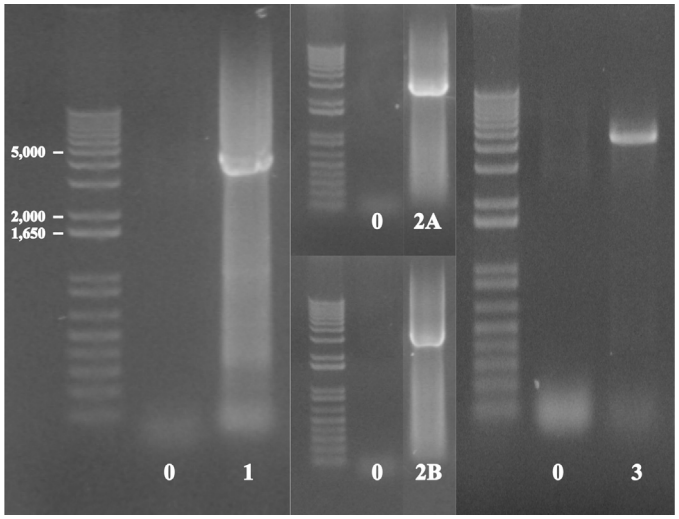


Fig. 1. Gel electrophoresis of PCR amplicons. A 1% agarose-gel showing the long-range PCR amplicons, fragment sizes are indicated with a 1 kb plus ladder. Lanes 0 represent negative control samples, lane 1 is AMDV-G amplified in one fragment (primers F1 + R3), lane 2A and 2B is AMDV-Utah amplified in two fragments; primers F1 + R2 and F2 + R3, respectively, and lane 3 is AMDV-G amplified in one fragment with primers F1 + R5.

industry (Decaro et al., 2012). The transmission patterns of AMDV between farms are not fully elucidated, and outbreak investigation is currently hampered by lack of sensitive tools for detection and typing of the virus. Previous studies have focused on smaller and conserved parts of the AMDV genome (Christensen et al., 2011; Knuuttila et al., 2015; Leimann et al., 2015; Oie et al., 1996) and therefore have produced data less suitable for typing. In addition, there are methods for characterisation of the AMDV genome using restriction fragmentation (Aasted, 1980) and Sanger sequencing (Alexandersen et al., 1988) but since they yield genetic information for limited stretches of the genome, they too provide less resolution than full genome sequencing. Next generation sequencing (NGS) is a powerful tool that has become cheaper and more easily available and it has successfully been applied to characterise entire genomes of other viruses and the genetic information obtained has been used to improve preventative measures (Escobar-Gutiérrez et al., 2012; Jakhesara et al., 2014; Kvisgaard et al., 2013).

To the authors' knowledge, the whole AMDV genome has not previously been sequenced using NGS. The aim of this study was to develop a fast, sensitive high-through-put method for full genome sequencing of the AMDV genome by NGS to lay the foundation for future development of tools for outbreak investigation, determination of virulence markers, and for development of more sensitive diagnostic tests and robust phylogenetic analyses.

2. Material and methods

2.1. Virus isolates

In order to establish an as universal method as possible two AMDV isolates with very different phenotypes, and hence presumably also different genotypes, were selected. The non-virulent cell-culture adapted strain AMDV-G (cell culture isolate, passage 10) was obtained from The Research Foundation of the Danish Fur Breeders' Association/Antigen Laboratory (Glostrup, DK), while the highly virulent AMDV-Utah isolate (antigen) was provided by emeritus Professor Bent Aasted (Copenhagen, DK). Total DNA was extracted using the QIAmp® MinElute Virus Spin Kit (Qiagen, Hilden, D) according to the manufacturer's instructions, and the final DNA elution was performed with 50 μL low TE-buffer.

Table 1
The primer sequences designed in the present study and the sizes of the expected amplicons for the applicable combinations. All primers have been designed with the Primer 3 software using the AMDV-G genome with accession number NC001662 as reference. Forward primers are indicated F, reverse primers R.

Primer	Positions (NC_001662)	Primer sequence (5′-3′)	Amplicon size (bp)
F1	77-97	CGCTTCGCGCTTGCTAACTTC	
R1	1814-1793	GCTCTGCGTGAGCGTTTGTTTC	
F1 + R1	77-1814		1735
F2	1502-1525	CCGGGGGGCACTGGAAAAACCTTG	
R2	3317-3296	GCAGAGAGGAGGTAGCCCCAAG	
F2 + R2	1502-2217		1816
F3	2934-2953	GCGTCGTTACAGGTTGCTTT	
R3	4467-4448	TTAATCCGCCACTTTCTGGT	
F3 + R3			1534
R5	4462-4439	CCGCCACTTTCTGGTAAAATAAGG	
F1 + R2			3240
F2 + R3			1946
F1 + R3			4390
F1 + R5			4385

2.2. DNA amplification

A long-range PCR covering about 91% of the AMDV genome (nucleotide position 98–4467, Table 1) was developed. The use of specific PCR amplification is important for future field applications as it avoids host DNA and attains a sufficient amount of double stranded DNA for the preparation of a sequencing library. The AMDV-genome was amplified in either a single or two overlapping fragments. The PCR primer-sequences are listed in Table 1 and were designed using the Primer3 software (Ye et al., 2012) with AMDV-G (accession no. NC001662) (Bloom et al., 1988) as reference genome. PCR reactions were setup as follows: 25 μL GoTag® Long PCR Master Mix (cat. no. M4021, Promega, Madison, WI), 2 μM primer F and R, 5 μL DNA template, and distilled water up to a total sample volume of 50 μL. Final cycling conditions were; initial denaturation at 95 °C for 2 min, 38 cycles of 30 s denaturation at 95 °C, 20 s annealing at 58 °C, and 30 s extension at 72 °C, followed by a final extension step for 10 min at 72 °C. All reactions were performed in a Bio-Rad CFX96 Touch instrument (Bio-Rad Laboratories, Inc., Hercules, CA).

The PCR products were analysed on 1% agarose gels stained with ethidiumbromide, and purified according to the manufacturer's instructions using the QIAquick® PCR Purification Kit or the QIAquick® Gel extraction kit (both from Qiagen, Hildren, D) depending on if there was a single product or not. DTU (Technical University of Denmark) Multi-Assay Core (Lyngby, Denmark) prepared the sequencing libraries according to the manufacturer's instructions and sequenced the samples on a 318-chip using the Ion Torrent PGM® (Life Technologies, Carlsbad, CA). The technology was chosen because it is easily accessible, cheap and fast, which are all-important factors for the future intended use in field applications. The region between positions 2470 and 2520 displayed very low read coverage in the Ion Torrent sequencing (possibly due to the nucleotide composition in this area) and we therefore Sanger-sequenced PCR-products spanning this region using primers F2 and R2 (Table 1) to verify the Ion-torrent findings.

2.3. Data analysis

Raw data in fastq-format were quality checked with FastQC version 0.10.1 and trimmed based on length (100–400 bp) and quality (average quality score >20) using Prinseq-lite (Schmieder and Edwards, 2011). Primer sequences were removed using Cutadapt version 1.4.1 (Martin, 2011). Reads were corrected for sequencing errors using RC454, and assembled with the associated Mosaik2

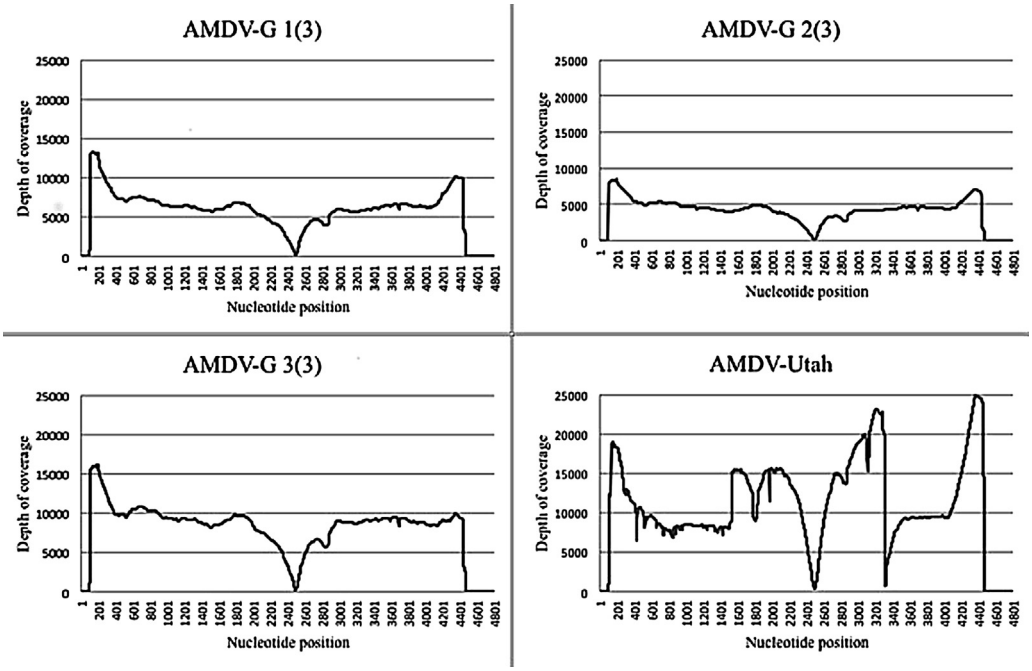


Fig. 2. Read coverage-plots. Coverage-plots showing the depth of coverage (*Y*-axis) at each sequenced nucleotide position in the genome (*X*-axis). The dip in coverage between position 2470 and 2520 is assumed to be caused by the ion-semiconductor sequencer having problems to read this homopolymeric region.

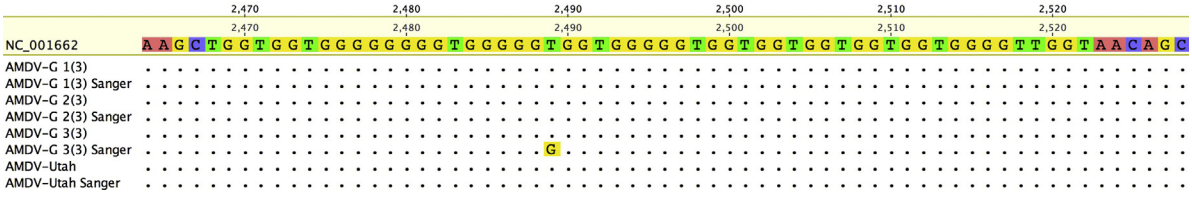


Fig. 3. Nucleotide alignment covering the homopolymeric region. Nucleotide alignment covering the homopolymeric region between positions 2470 and 2520 (repeated stretches of G's). Consensus sequences for each triplicate of AMDV-G, for AMDV-Utah, and for their corresponding Sanger generated sequence. The single nucleotide difference at position 2489 in the Sanger sequence for the third replicate of AMDV-G (3(3)) is highlighted.

assembler (Henn et al., 2012). The error-corrected reads were then mapped using BWA (Li, 2013) to the AMDV-G reference (NC001662) cut between the 5′ and 3′ annealing sites for primers F1 and R3 respectively (Table 1). For each sample a consensus sequence was generated using Vcf-tools 0.1.12a (Danecek et al., 2011). Multiple alignments of the full genomes was done using MAFFT v7.205 (Katoh and Standley, 2013). For additional comparisons, the following AMDV-Utah sequences were downloaded from the NCBI database: U39015.1, X77083.1 and Z1827.6.1. Alignments were visualized in Geneious 7.1.5. (Kearse et al., 2012).

3. Results

3.1. Specificity of the PCR products

A long range PCR assay for specific amplification of the AMDV genome was developed. Originally, three sets of primers spanning the AMDV genome were designed and the non-coding palindromic 5′- and 3′-ends (Bloom et al., 1990) which are known to interfere with PCR-amplification were excluded. After running a matrix with the possible combinations of forward and reverse primers, the optimal primer pairs was selected as assessed by gel electrophoresis. PCR cycling conditions were optimized by running an annealing-temperature gradient and amplification of PCR-products of the expected sizes were confirmed by gel electrophoresis (Fig. 1). The PCR reactions that produced one specific product were directly processed for sequencing. In cases where additional PCR-products were present, the band of the expected size was extracted from the gel prior to sequencing (Fig. 1). The AMDV-G and AMDV-Utah sequences sequenced in this study was amplified using primer F1 + R3 (Table 1), however during further assay optimisation better yields of PCR-product was achieved using primer F1 and R5 (assed by gel electrophoresis, Fig. 1). Despite the presumed genetic differences between the two viral isolates they both amplified well using these primers (Fig. 1).

3.2. Sequence quality and coverage

AMDV-G DNA was extracted in triplicates, and each individual sample was PCR-amplified and sequenced. AMDV-Utah DNA was extracted once, PCR-amplified, and sequenced. Primer-sequences and low quality reads were removed prior to data analysis. The data quality was overall high, and for each sample approximately 99% of the trimmed and quality filtered reads mapped to the AMDV-G reference. Coverage-plots for each of the four samples showed a dip in coverage between nucleotide positions 2470 and 2520 (Fig. 2). However, the Sanger-generated sequences spanning this region matched the sequences produced by the Ion Torrent, with the exception of a single nucleotide difference in the Sanger

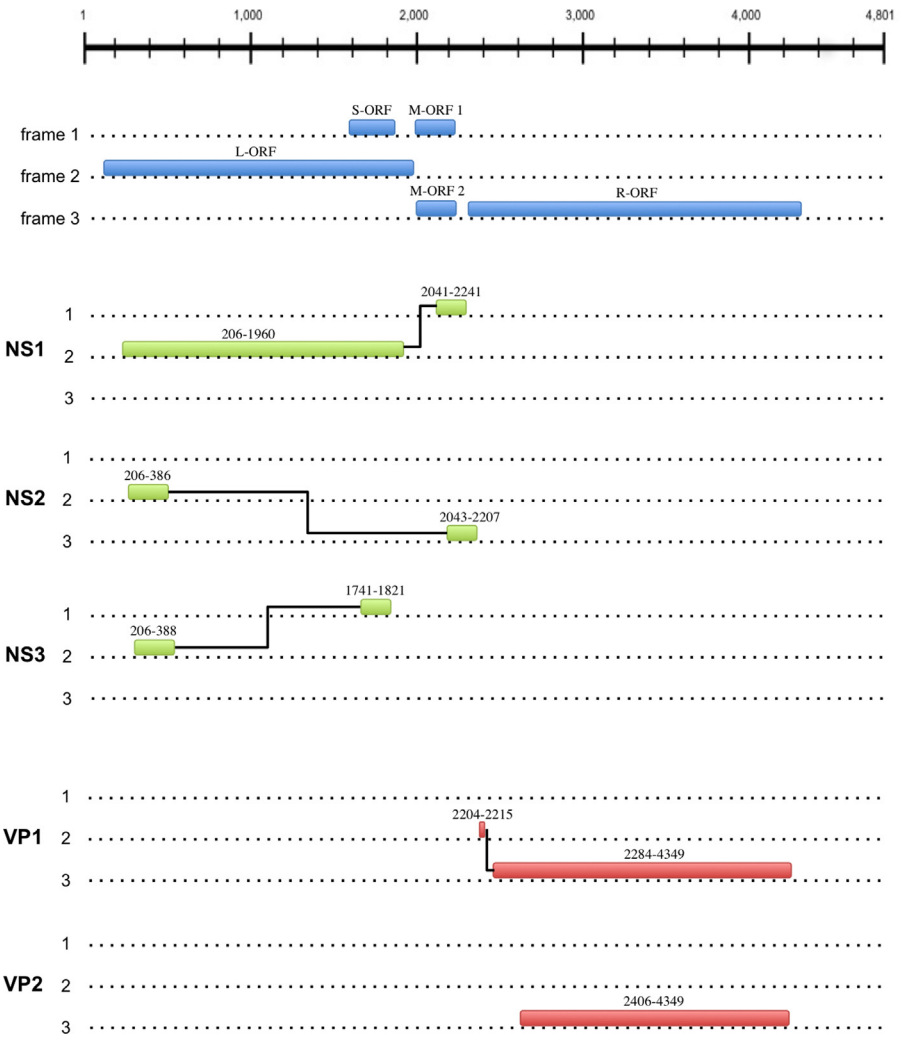


Fig. 4. Genomic map of AMDV. Genomic map of AMDV showing the open reading frames (ORF's) and how the major proteins are spliced together. Nucleotide positions refer to the reference AMDV-G genome. Nucleotide positions: 116–1975 left ORF (L-ORF), 1535–1825 small ORF (S-ORF), 1993–2209 mid ORF 1 (M-ORF 1), 1983–2204 mid ORF 2 (M-ORF 2), and 2241–4346 right ORF (R-ORF).

sequence for AMDV-G replicate no. 3 (Fig. 3). The quality score for this deviating base (G) was low (eight), while the scores for the corresponding base in the other two Sanger sequences were 38 and 40, and therefore indicates an error from the Sanger sequencing. The AMDV-G and AMDV-Utah sequences have the following accession numbers: KU513985, KU513986, KU513987, and KU513988, respectively.

3.3. Sequence analysis

The alignments showed an overall high sequence similarity (99% homology) between the AMDV-G samples generated in this study and the previously published AMDV-G sequence obtained by Sanger sequencing (NC001662). The robustness of the protocol was demonstrated by aligning the sequences obtained from sequencing AMDV-G in triplicates, and a deviation of a single nucleotide was found, out of the total 3369. An overview of the AMDV genomic organization is provided in Fig. 4. The AMDV-G triplicates and AMDV-Utah were compared on nucleotide and amino acid level (Table 2), and unless other is indicated, the results and discussion compare the AMDV-G and AMDV-Utah strains generated in the present study.

3.4. NS1 gene

The left open reading frame encodes for NS1–3. NS1 is the major regulatory protein in parvoviruses and it plays an important role in viral replication during infection (Fields et al., 2007; Gottschalck et al., 1994). In addition to confirming a number of previously reported differences between AMDV-G and AMDV-Utah, the present study also report novel findings as demonstrated in Table 2.

In the purine binding pocket between amino acids 421–492 (Gottschalck et al., 1994) a single change (F481L) between AMDV-G and AMDV-Utah was observed in addition to a single change (F430L) between ADMV-G and the AMDV-G reference (Table 2). Overall, the purine binding region, including the GKRN-region between amino acids 435–440 and its purine binding pocket, was well conserved in the sequences produced in the present study (Fig. 5, panel A). In agreement with previous studies of the distribution of changes in NS1 (Gottschalck et al., 1994), a higher degree of variability was demonstrated in the N- and C-terminals of NS1 compared to in the middle (Fig. 5 panel A).

Table 2
Overview of the nucleotide (nt) changes reported in the present study, including changes in the affected amino acids (aa) for the two most well described genes; NS1 and VP2. Absolute nucleotide positions refer to the AMDV-G reference genome (NC001662). Dash/es means no change in comparison to the AMDV-G reference genome. Isolates sequenced in the present study are indicated by *, and observations without a reference originates from the present study and hence are novel.

Absolute nt pos.	Codon pos.	NC001662		AMDV-Utah*		AMDV-G*		Reference accessions
		nt	aa	nt	aa	nt	aa	
L-ORF								
179–181		AGC	S	-	-	G--	G	
221–223	NS1/2 = 6	ATT	I	C--	L	---	-	Z18276.1, X77083.1
275–277	NS1/2 = 24	AAC	N	GCT	A	---	-	
290–292	NS1/2 = 29	GTT	V	C--	L	---	-	
293–295	NS1/2 = 30	GCC	A	--T	-	---	-	
299–301	NS1/2 = 32	TTG	L	C-A	-	---	-	
353–355	NS1/2 = 50	CCG	P	--A	-	---	-	Z18276.1, X77083.1
368–370	NS1/2 = 55	ACC	T	--T	-	---	-	
389–391	NS1 = 62	GCT	A	--A	-	---	-	
395–297	NS1 = 64	GAC	D	--T	-	---	-	
410–412	NS1 = 69	AAT	N	-CC	T	---	-	
416–418	NS1 = 71	ACA	T	-T-	I	---	-	
431–433	NS1 = 76	CAC	H	--G	Q	---	-	Z18276.1, X77083.1
440–442	NS1 = 79	AAC	N	--A	K	---	-	
443–445	NS1 = 80	AAT	N	G--	D	---	-	
470–472	NS1 = 89	TTG	L	--A	-	---	-	
485–487	NS1 = 94	CTG	L	G--	V	---	-	
491–493	NS1 = 96	ATT	I	G--	V	---	-	C--/L Z18276.1, X77083.1
506–508	NS1 = 101	AAA	K	--G	-	---	-	
509–511	NS1 = 102	AGC	S	--T	-	---	-	
524–526	NS1 = 107	AGT	S	GC-	A	---	-	-A-/N Z18276.1, X77083.1
527–529	NS1 = 108	AAC	N	G-T	D	---	-	
533–535	NS1 = 110	GTT	V	A--	I	---	-	
539–541	NS1 = 112	TTA	L	--C	F	---	-	
542–544	NS1 = 113	ATT	I	--C	-	---	-	
572–574	NS1 = 123	CAA	Q	--C	H	---	-	
650–652	NS1 = 149	TTT	F	--G	L	---	-	
653–655	NS1 = 150	ATG	M	--T	I	---	-	
659–661	NS1 = 152	AGA	R	-A-	K	---	-	
668–670	NS1 = 155	AAA	K	-G-	R	---	-	
680–682	NS1 = 159	GTT	V	-C-	A	---	-	C--/L Z18276.1, X77083.1
686–688	NS1 = 161	TAT	Y	-A-	F	---	-	
707–709	NS1 = 168	ATA	I	CA-	Q	---	-	
713–715	NS1 = 170	GAT	D	--C	-	---	-	
728–730	NS1 = 175	GAA	E	--G	-	---	-	
731–733	NS1 = 176	GAT	D	-CC	A	---	-	
734–736	NS1 = 177	AGA	R	-A-	K	---	-	
740–742	NS1 = 179	AAG	K	--T	N	---	-	
746–748	NS1 = 181	CTA	L	T-G	-	---	-	
767–769	NS1 = 188	GGA	G	--G	-	---	-	
776–778	NS1 = 191	AAG	K	--A	-	---	-	
788–790	NS1 = 195	TAT	Y	--C	-	---	-	
791–793	NS1 = 196	TTT	F	-A-	Y	---	-	
818–829	NS1 = 205	AAT	N	--C	-	---	-	Z18276.1, X77083.1
830–832	NS1 = 209	CAC	H	AC-	T	---	-	
836–838	NS1 = 211	AGA	R	--T	S	---	-	
845–847	NS1 = 214	ACA	T	GT-	V	---	-	
848–859	NS1 = 2015	TTC	F	A-A	I	---	-	
878–890	NS1 = 225	AAT	N	C--	H	-		
881–883	NS1 = 226	ACA	T	-AG	K	-	-	
884–886	NS1 = 227	GAT	D	--A	E	-	-	
887–889	NS1 = 228	AGT	S	G--	G	-	-	
902–904	NS1 = 233	TTT	F	-A-	Y	-	-	
920–922	NS1 = 239	GGC	G	--T	-	-	-	
923–925	NS1 = 240	ATT	I	--C	-	-	-	
925–928	NS1 = 241	GTT	V	A--	I	-	-	
953–955	NS1 = 250	AAA	K	--G	-	-	-	
954–056	NS1 = 251	ACT	T	G-C	A	-	-	
974–976	NS1 = 257	TTA	L	--G	-	-	-	
980–982	NS1 = 259	GAG	E	--A	-	-	-	
1007–1009	NS1 = 268	AAT	N	G--	D	-	-	
1025–1027	NS1 = 274	GGC	G	---	-	A--	S	
1070–1072	NS1 = 289	ACA	T	T--	S	-	-	
1106–1108	NS1 = 301	GCT	A	--A	-	---	-	
1109–1111	NS1 = 302	ACT	T	--C	-	---	-	
1130–1132	NS1 = 309	GAA	E	---	-	--C	D	
1154–1156	NS1 = 317	CCT	P	---	-	--G	-	
1181–1183	NS1 = 326	AGT	S	-A-	N	---	-	Z18276.1, X77083.1
1337–1339	NS1 = 378	ATT	I	--G	M	---	-	Z18276.1, X77083.1
1493–1495	NS1 = 430	TTC	F	---	-	C--	L	
1646–1648	NS1 = 481	TTT	F	C--	L	---	-	Z18276.1, X77083.1
1703–1705	NS1 = 500	GAC	D	--T	-	---	-	Z18276.1, X77083.1

Table 2 (*Continued*)

Absolute nt pos.	Codon pos.	NC001662		AMDV-Utah*		AMDV-G*		Reference accessions
		nt	aa	nt	aa	nt	aa	
2143–2145	NS1 = 620	TGC	C	---	-	-A-	Y	
2161–2163	NS1 = 626	AGT	S	G--	G	---	-	Z18276.1, X77083.1
M-ORF1								
2143–2145	NS1 = 620	TGC	C	---	-	-A-	Y	
2161–2163	NS1 = 626	AGT	S	G--	G	-	-	Z18276.1, X77083.1
2218–2220	non-coding	ATA	I	G--	V	G--	V	2179: GCA/A to CCG/P Z18276.1, X77083.1 Z18276.1, X77083.1
M-ORF2								
2142–2144	NS2 = 94	CTG	L	---	-	--A	-	
2160–2162	NS2 = 100	GAG	E	-G-	G	---	-	Z18276.1, X77083.1
2217–2219	non-coding	AAT	N	-G-	S	-G-	S	Z18276.1, X77083.1
R-ORF								
2631–2633	VP2 = 76	GAC	D	--T	-	---	-	Z18276.1, U39015.1
2673–2675	VP2 = 90	AAA	K	---	-	C--	Q	
2679–2681	VP2 = 92	CAT	H	---	-	---	H	GC-/A Z18276.1, U39015.1
2685–2687	VP2 = 94	CAA	Q	---	-	---	-	A--/K Z18276.1, U39015.1
2748–2750	VP2 = 115	TAT	Y	-T-	F	---	-	Z18276.1, U39015.1
2751–2753	VP2 = 116	ATA	I	T--	L	---	-	
3459–3461	VP2 = 352	ATT	I	G--	V	---	-	Z18276.1, U39015.1
3585–3587	VP2 = 394	CAA	Q	---	-	--G	-	
3588–3590	VP2 = 395	CAC	H	--G	Q	A--	N	Z18276.1, U39015.1
3693–3695	VP2 = 430	TAC	Y	---	-	--T	-	
3696–3698	VP2 = 431	TAC	Y	---	-	ATT	I	
3705–3707	VP2 = 434	AAT	N	CAT	H	---	-	Z18276.1, U39015.1
3876–3878	VP2 = 491	AAC	N	G--	D	---	-	GAG/E Z18276.1, U39015.1
3975–3977	VP2 = 524	CCG	P	--A	-	--A	-	Z18276.1, U39015.1
4005–4007	VP2 = 534	CAT	H	G--	D	---	-	Z18276.1, U39015.1
4125–4127	VP2 = 574	AAT	N	---	-	G--	D	
4263–4265	VP2 = 620	AAG	K	AAC	N	---	-	Z18276.1, U39015.1
4305–4307	VP2 = 634	ATA	I	---	-	--G	M	

3.5. VP2 gene

In addition to confirming a number of previously reported differences between AMDV-G and AMDV-Utah, some of which have been proposed to influence virulence and host range (overview in Table 2), this study report novel differences in the VP2 gene. The N-terminus of VP2, amino acid 1–220, has been suggested to play a role in AMDV host range and culturing ability (Bloom et al., 1998), and the present study confirm some, but not all, of the previously reported differences between AMDV-G and AMDV-Utah in this region (Table 2). In addition a novel change in AMDV-Utah (T116L) is reported here.

Amino acid 420 have been proposed to increase viral fitness by prevention of caspase cleavage (Cheng et al., 2010), however in agreement with other studies (Bloom et al., 1988; Oie et al., 1996; Sang et al., 2012) that particular difference between AMDV-G and AMDV-Utah was not observed here either. VP2 amino acid 428–446 functions as a small part of the capsid which has also been suggested to be important for immunopathogenesis by defining AMDV host range (McKenna et al., 1999). The present study confirms a previously reported difference, N343H, in this area, but whether or not this change results in increased pathogenicity is currently unknown (Bloom et al., 1988; Oie et al., 1996; Sang et al., 2012).

In addition to the above-mentioned differences between AMDV-G and AMDV-Utah, additional differences were observed in the VP2 amino acid sequence between the AMDV-G reference and the AMDV-G strain from the laboratory (e.g. K90Q, H395N and D574N). It is currently unknown if there is a fitness effect associated to these changes (e.g. adaption to tissue-culture conditions).

The overall lower conservation of AMDV NS-genes compared to other parvoviruses is supported in the present study by the higher degree of variability in the left ORF compared to in the right ORF, which was even more striking on amino acid level (Fig. 5, panel A).

3.6. Regulatory elements

Previous studies have identified eight TATA-boxes in the AMDV genome; two confirmed functional at nucleotide 154 (TATAA) and 1729 (TATTAA), and six additional boxes at nucleotide 665, 818, 2546, 4136, 4394, and 4468 (AATAAA) with unknown function (Bloom et al., 1988). In the AMDV-Utah sequence, a previously reported difference to AMDV-G in TATA-box 818 (T820C) (Bloom et al., 1988; Gottschalck et al., 1994) is confirmed, and a previously not described change in the 665 box (A669G) is reported. But since the function of these TATA boxes is unknown, the importance of these differences remains to be investigated. The 4468 box was not included in the sequences generated in the present study.

A P3 promotor that initiates transcription of all mRNA have previously identified at nucleotide 151–160 (p3, GTATATAAGC) (Bloom et al., 1988), in addition to a more uncertain promotor P36 around nucleotide 1744 (Bloom et al., 1988; Qiu et al., 2006). In the present study these promotors were fully conserved. However at the suggested transcription initiation site at nucleotide 179 a change (A179G) is reported in the present AMDV-G strain (Table 2) compared to the reference AMDV-G genome.

Internal polyadenylation sites (pA)p's at nucleotide 2561 and 4391 have been suggested to play a major role in AMDV replication (Huang et al., 2012). Studies in the parvovirus Minute Virus of Mice (MVM) shows that the above mentioned NS1 GKRN-region contains a NS1 recognition site at amino acids 337–344 (ACCAACCA), which together with an upstream nicking site initiates viral replication (Christensen et al., 1997). All of these sites were conserved in the two viruses investigated here, and therefore should not have any effect on the increased virulence of AMDV-Utah.

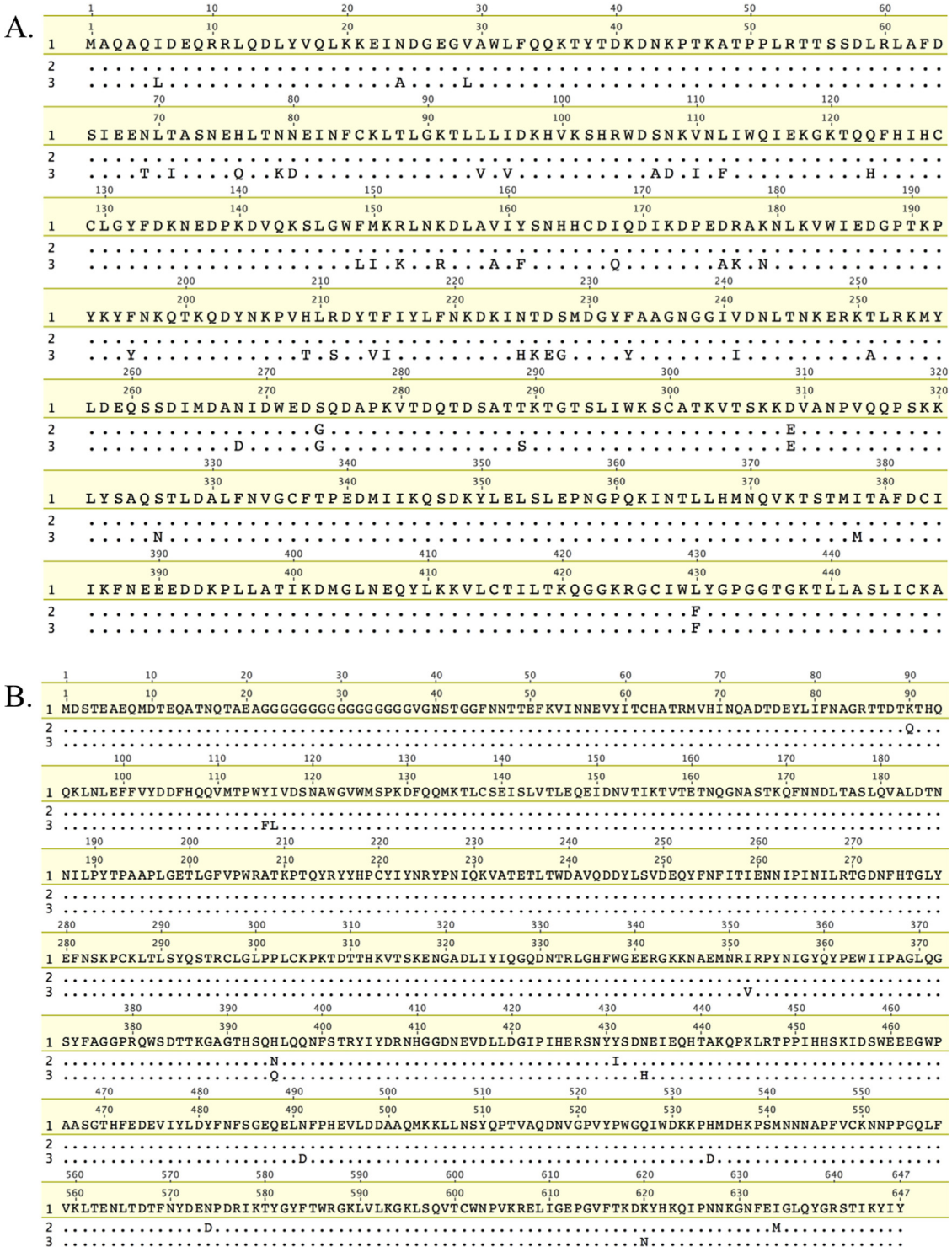


Fig. 5. Protein alignments for each of the two major AMDV genes. Translation and alignments at protein level for each of the NS1 (Panel A) and VP2 (Panel B) genes for the AMDV-G reference NC_001662 (1), one representative AMDV-G strain (2), and the AMDV-Utah (3) strain sequenced in this study.

4. Discussion

This paper describes a fast and robust protocol for next generation sequencing of the near full length AMDV genome and the subsequent data analysis. The protocol was verified by gel electrophoresis, complimentary Sanger sequencing, and by sequence analysis. The prototypic non-virulent cell-culture adapted strain AMDV-G was used as a model virus, and for comparison and to investigate genetic virulence markers the highly virulent AMDV-Utah strain was also sequenced. Due to the presence of secondary structures and palindromic motifs at the 3′ and 5′ ends, approximately 91% of the viral genome was amplified; including all known coding regions (Alexandersen et al., 1988; Bloom et al., 1988).

This study confirm some of the previously reported nucleotide and amino acid differences between AMDV-G and AMDV-Utah (Table 2) and no major deviations in the suggested genomic regulatory regions were observed (Bloom et al., 1988). Therefore, one can speculate that the increased virulence of AMDV-Utah compared

to AMDV-G is not due to differences in gene regulation but rather on protein level.

However, some nucleotide and amino acid differences was observed between the previously published Sanger generated AMDV-G and AMDV-Utah genomes and the genomes sequenced in this study, which could be a result of further cell-culture adaptation of both strains or due to the use of different sequencing technologies. One specific change of interest is the A179 G seen in our AMDV-G strain (Table 2), as it might influence the translation initiation codon at nt 179–181. Another change that might be of importance is the F430L. It resides in close proximity to the conserved ATP-binding pocket and GKRN region between amino acids 435–440, which has been suggested to be essential for the NS1 protein and viral DNA replication due to its ATP- and GTP binding sites and its ATPase function (Gottschalck et al., 1994).

There is some disagreement in the literature regarding the frame and sequence of the first three amino acids of the AMDV-G VP1 gene, in that one study suggests the start to be MSK in frame 2 (Huang et al., 2012) (accession number JN040434.1), while another study suggests HHN in frame 3 (Schuierer et al., 1997) (accession number X97629.1). The first option would be more similar to e.g. human parvovirus B19, in which the VP1-unique region (starting with an MSK) is encoded in another frame, while the remaining protein is identical to that of VP2 (as depicted in Fig. 4). In the second option, starting with amino acids HHN, the whole VP1 protein is encoded in the same frame as that of VP2, but with an additional 55 amino acids in the N-terminal. However, in this sequence there is no start codon (Fig. 4), and therefore it is sensible to assume that option one (MSK) is more correct. In human parvovirus B19 this unique VP1 N-terminal has been identified as key for viral entry (Leisi et al., 2013), and studies in AMDV suggest its importance as it provides phospholipase A2 enzyme activity which modifies the endosome membrane thereby mediating capsid release (Fenner's Veterinary Virology, 2011). It has further been suggested that AMDV's ability to grow in cell culture is regulated by the N-terminal of the VP2 gene (Bloom et al., 1998), and one can therefore speculate that the VP1 unique terminal could be linked to *in vivo* infectivity.

Both the VP and the NS proteins have been suggested to be involved in determining the viral host range and influence pathogenicity (Fields et al., 2007). For example, it has been shown that knockout of the NS1-gene resulted in failure to produce replicative form AMDV DNA (Huang et al., 2014). In the present study, the majority of nucleotide and amino-acid differences between AMDV-G and AMDV-Utah were in the left ORF. This is in agreement with previous studies reporting that the non-structural (NS) proteins of AMDV-G and AMDV-Utah have different molecular weights (Alexandersen et al., 1986), and are less conserved than in other parvoviruses (Gottschalck et al., 1994). Interestingly, it is known that the AMDV right ORF is more conserved despite containing virulence factors important for the viral entry (Gottschalck et al., 1991; Oie et al., 1996). These findings indicate that the virulence of AMDV-Utah may not be primarily due to increased infectivity since this function depends primarily on the VP genes. Instead the difference in virulence could e.g. be linked to the NS proteins that are involved with virion assembly, release, unpacking, or the ability to avoid host cell responses.

The protocols developed in this study enable viral DNA to be extracted and amplified from primary sample material and by that avoiding the use of labour intensive cloning to amplify the viral DNA prior to sequencing. The PCR-amplification step is also useful as the concentration of viral DNA in viraemic animals is not sufficient to directly act as template for next generation sequencing. The viral strains used to establish this protocol have very different phenotypes and despite the expected genetic difference both strains amplified well. This indicates a high degree of conservation in the primer-annealing region, which is further supported by on-going work successfully amplifying AMDV field strains (data not shown). There are however potential biases when using PCR amplified DNA as input for sequencing, e.g. for investigating quasi-species. But this is of less importance if the resulting data will be used for comparing sample consensus sequences, as in the present study. The ion-semiconductor technology is known to have difficulties to accurately PCR amplify and read homopolymers (Quail et al., 2012), and especially G'- and C'-rich regions as between position 2470 and 2520 in the AMDV-G reference genome. Therefore the robustness of the protocols was demonstrated by processing the AMDV-G strain in triplicates and by sequencing the homopolymeric region in each sample using the complimentary Sanger sequencing method. Since the NGS sequences from each sample were identical, and so were the Sanger generated fragments (the latter with the exception of one erroneous base), it was concluded that the dip in coverage was caused during the sequencing, and not by the PCR amplification. Thus, it would be beneficial to disregard this region in alignments when the Ion Torrent technology is used, and for the development of molecular tools.

In conclusion, this is to the authors' knowledge the first study to describe the entire coding sequence of the AMDV genome using next generation sequencing. The study provides a robust and fast method for generating whole genome sequences of AMDV from various DNA sources and will create value by allowing for phylogenetic analysis with higher resolution and by facilitating development of new diagnostic tools.

Acknowledgements

Kopenhagen Diagnostics, Kopenhagen Fur, and Professor emeritus Bent Aasted, UCPH, are gratefully thanked for supplying the viral strains AMDV-G and AMDV-Utah, respectively.

References

Aasted, B., 1980. Purification and characterization of Aleutian disease virus. Acta Pathol. Microbiol. Scand. B 80, 323–328.

Alexandersen, S., Uttenthal-Jensen, A., Aasted, B., 1986. Demonstration of non-degraded aleutian disease virus (ADV) proteins in lung tissue from experimentally infected mink kits. Arch. Virol. 87, 127–133, http://dx.doi.org/10.1007/BF01310549.

Alexandersen, S., Bloom, M.E., Perryman, S., 1988. Detailed transcription map of Aleutian mink disease parvovirus. J. Virol. 62, 3684–3694.

Bloom, M.E., Race, R.E., Wolfinbarger, J.B., 1980. Characterization of Aleutian disease virus as a parvovirus. J. Virol. 35, 836–843.

Bloom, M.E., Alexandersen, S., Perryman, S., Lechner, D., Wolfinbarger, J.B., 1988. Nucleotide sequence and genomic organization of Aleutian mink disease parvovirus (ADV): sequence comparisons between a nonpathogenic and a pathogenic strain of ADV. J. Virol. 62, 2903–2915.

Bloom, M.E., Alexandersen, S., Garon, C.F., Mori, S., Wei, W., Perryman, S., Wolfinbarger, J.B., 1990. Nucleotide sequence of the 5′-terminal palindrome of Aleutian mink disease parvovirus and construction of an infectious molecular clone. J. Virol. 64, 3551–3556.

Bloom, M.E., Fox, J.M., Berry, B.D., Oie, K.L., Wolfinbarger, J.B., 1998. Construction of pathogenic molecular clones of Aleutian mink disease parvovirus that replicate both in vivo and in vitro. Virology 251, 288–296, http://dx.doi.org/10.1006/viro.1998.9426.

Broll, S., Alexandersen, S., 1996. Investigation of the pathogenesis of transplacental transmission of Aleutian mink disease parvovirus in experimentally infected mink. J. Virol. 70, 1455–1466.

Cheng, F., Chen, A.Y., Best, S.M., Bloom, M.E., Pintel, D., Qiu, J., 2010. The capsid proteins of Aleutian mink disease virus activate caspases and are specifically cleaved during infection. J. Virol. 84, 2687–2696, http://dx.doi.org/10.1128/JVI.01917-09.

Christensen, J., Cotmore, S.F., Tattersall, P., 1997. Parvovirus initiation factor PIF: a novel human DNA-binding factor which coordinately recognizes two ACGT motifs. J. Virol. 71, 5733–5741.

Christensen, L.S., Gram-Hansen, L., Chriél, M., Jensen, T.H., 2011. Diversity and stability of Aleutian mink disease virus during bottleneck transitions resulting from eradication in domestic mink in Denmark. Vet. Microbiol. 149, 64–71, http://dx.doi.org/10.1016/j.vetmic.2010.10.016.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., McVean, G., Durbin, R., 2011. The variant call format and VCFtools. Bioinformatics 27, 2156–2158, http://dx.doi.org/10.1093/bioinformatics/btr330.

Danish Executive Order 1447 of 15/12/2009, 2009, Danish Executive Order 1447 of 15/12/2009.
Decaro, Nicola, Buonavoglia, C., Ryser-Degiorgis, M.-P., Gortázar, C., 2012. 12. Parvovirus infections. In: Gavier-Widén, D., Duff, J.P., Meredith, A. (Eds.), Infectious Diseases of Wild Mammals and Birds in Europe. Wiley-Blackwell Oxford, UK, pp. 181–285.
Escobar-Gutiérrez, A., Vazquez-Pichardo, M., Cruz-Rivera, M., Rivera-Osorio, P., Carpio-Pedroza, J.C., Ruíz-Pacheco, J.A., Ruiz-Tovar, K., Vaughan, G., 2012. Identification of hepatitis C virus transmission using a next-generation sequencing approach. J. Clin. Microbiol. 50, 1461–1463, http://dx.doi.org/10.1128/JCM.00005-12.
Fenner's Veterinary Virology, 2011. Fenner's Veterinary Virology. Elsevier, http://dx.doi.org/10.1016/B978-0-12-375158-4.00012-2.
Fields, B.N., Knipe, D.M., Howley, P.M., 2007. Fields Virology, 5th ed. Fields Virol.
Gottschalck, E., Alexandersen, S., Cohn, A., Poulsen, L.A., 1991. Nucleotide sequence analysis of Aleutian mink disease parvovirus shows that multiple virus types are present in infected mink. J. Virol. 65, 4378–4386.
Gottschalck, E., Alexandersen, S., Storgaard, T., Bloom, M.E., Aasted, B., 1994. Sequence comparison of the non-structural genes of four different types of Aleutian mink disease parvovirus indicates an unusual degree of variability. Arch. Virol. 138, 213–231, http://dx.doi.org/10.1007/BF01379127.
Henn, M.R., Boutwell, C.L., Charlebois, P., Lennon, N.J., Power, K.A., Macalalad, A.R., Berlin, A.M., Malboeuf, C.M., Ryan, E.M., Gnerre, S., Zody, M.C., Erlich, R.L., Green, L.M., Berical, A., Wang, Y., Casali, M., Streeck, H., Bloom, A.K., Dudek, T., Tully, D., Newman, R., Axten, K.L., Gladden, A.D., Battis, L., Kemper, M., Zeng, Q., Shea, T.P., Gujja, S., Zedlack, C., Gasser, O., Brander, C., Hess, C., Günthard, H.F., Brumme, Z.L., Brumme, C.J., Bazner, S., Rychert, J., Tinsley, J.P., Mayer, K.H., Rosenberg, E., Pereyra, F., Levin, J.Z., Young, S.K., Jessen, H., Altfeld, M., Birren, B.W., Walker, B.D., Allen, T.M., 2012. Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. PLoS Pathog. 8, e1002529, http://dx.doi.org/10.1371/journal.ppat.1002529.
Huang, Q., Deng, X., Best, S.M., Bloom, M.E., Li, Y., Qiu, J., 2012. Internal polyadenylation of parvoviral precursor mRNA limits progeny virus production. Virology 426, 167–177, http://dx.doi.org/10.1016/j.virol.2012.01.031.
Huang, Q., Luo, Y., Cheng, F., Best, S.M., Bloom, M.E., Qiu, J., 2014. Molecular characterization of the small nonstructural proteins of parvovirus Aleutian mink disease virus (AMDV) during infection. Virology 452–453, 23–31, http://dx.doi.org/10.1016/j.virol.2014.01.005.
Jakhesara, S.J., Bhatt, V.D., Patel, N.V., Prajapati, K.S., Joshi, C.G., 2014. Isolation and characterization of H9N2 influenza virus isolates from poultry respiratory disease outbreak. Springerplus 3, 196, http://dx.doi.org/10.1186/2193-1801-3-196.
Katoh, K., Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30, 772–780, http://dx.doi.org/10.1093/molbev/mst010.
Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., Drummond, A., 2012. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28, 1647–1649, http://dx.doi.org/10.1093/bioinformatics/bts199.
Knuuttila, A., Aaltonen, K., Virtala, A.-M.K., Henttonen, H., Isomursu, M., Leimann, A., Maran, T., Saarma, U., Timonen, P., Vapalahti, O., Sironen, T., 2015. Aleutian mink disease virus in free-ranging mustelids in Finland—a cross-sectional epidemiologic and phylogenetic study. J. Gen. Virol., http://dx.doi.org/10.1099/vir.0.000081.
Kvisgaard, L.K., Hjulsager, C.K., Fahnoe, U., Breum, S.O., Ait-Ali, T., Larsen, L.E., 2013. A fast and robust method for full genome sequencing of porcine reproductive and respiratory syndrome virus (PRRSV) type 1 and type 2. J. Virol. Methods 193, 697–705, http://dx.doi.org/10.1016/j.jviromet.2013.07.019.
Leimann, A., Knuuttila, A., Maran, T., Vapalahti, O., Saarma, U., 2015. Molecular epidemiology of Aleutian mink disease virus (AMDV) in Estonia, and a global phylogeny of AMDV. Virus Res. 199, 55–61, http://dx.doi.org/10.1016/j.virusres.2015.01.011.
Leisi, R., Ruprecht, N., Kempf, C., Ros, C., 2013. Parvovirus B19 uptake is a highly selective process controlled by VP1u, a novel determinant of viral tropism. J. Virol. 87, 13161–13167, http://dx.doi.org/10.1128/JVI.02548-13.
Li, H., 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.
Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 17, 10–12, http://dx.doi.org/10.14806/ej.17.1.200.
McKenna, R., Olson, N.H., Chipman, P.R., Baker, T.S., Booth, T.F., Christensen, J., Aasted, B., Fox, J.M., Bloom, M.E., Wolfinbarger, J.B., Agbandje-McKenna, M., 1999. Three-dimensional structure of Aleutian mink disease parvovirus: implications for disease pathogenicity. J. Virol. 73, 6882–6891.
Oie, K.L., Durrant, G., Wolfinbarger, J.B., Martin, D., Costello, F., Perryman, S., Hogan, D., Hadlow, W.J., Bloom, M.E., 1996. The relationship between capsid protein (VP2) sequence and pathogenicity of Aleutian mink disease parvovirus (ADV): a possible role for raccoons in the transmission of ADV infections. J. Virol. 70, 852–861.
Qiu, J., Cheng, F., Burger, L.R., Pintel, D., 2006. The transcription profile of Aleutian mink disease virus in CRFK cells is generated by alternative processing of pre-mRNAs produced from a single promoter. J. Virol. 80, 654–662, http://dx.doi.org/10.1128/JVI.80.2.654-662.2006.
Quail, M.A., Smith, M., Coupland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A., Swerdlow, H.P., Gu, Y., 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, pacific biosciences and Illumina MiSeq sequencers. BMC Genomics 13, 341, http://dx.doi.org/10.1186/1471-2164-13-341.
Sang, Y., Ma, J., Hou, Z., Zhang, Y., 2012. Phylogenetic analysis of the VP2 gene of Aleutian mink disease parvoviruses isolated from 2009 to 2011 in China. Virus Genes 45, 31–37, http://dx.doi.org/10.1007/s11262-012-0734-9.
Schmieder, R., Edwards, R., 2011. Quality control and preprocessing of metagenomic datasets. Bioinformatics (Oxford, England) 27, 863–864, http://dx.doi.org/10.1093/bioinformatics/btr026.
Schuierer, S., Bloom, M.E., Kaaden, O.R., Truyen, U., Diseases, E., Disease, I., 1997. Sequence analysis of the lymphotropic Aleutian disease parvovirus ADV-SL3. Brief Report, 157–166.
Shackelton, L.A., Hoelzer, K., Parrish, C.R., Holmes, E.C., 2007. Comparative analysis reveals frequent recombination in the parvoviruses. J. Gen. Virol. 88, 3294–3301, http://dx.doi.org/10.1099/vir.0.83255-0.
Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S., Madden, T.L., 2012. Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. BMC Bioinf. 13, 134, http://dx.doi.org/10.1186/1471-2105-13-134.

4.2. MANUSCRIPT 2

Evolutionary analysis of whole genome sequences from Aleutian Mink Disease Viruses confirms inter-farm transmission.

Status: accepted for publication in Journal of General Virology (JGV), March 2017.

(page numbers are relative to paper)

JOURNAL OF GENERAL VIROLOGY

RESEARCH ARTICLE
Hagberg *et al.*, *Journal of General Virology*
DOI 10.1099/jgv.0.000777

MICROBIOLOGY SOCIETY

Evolutionary analysis of whole-genome sequences confirms inter-farm transmission of Aleutian mink disease virus

Emma E. Hagberg,[1,2,*†] Anders G. Pedersen,[2] Lars E. Larsen[3] and Anders Krarup[1]

Abstract

Aleutian mink disease virus (AMDV) is a frequently encountered pathogen associated with mink farming. Previous phylogenetic analyses of AMDV have been based on shorter and more conserved parts of the genome, e.g. the partial NS1 gene. Such fragments are suitable for detection but are less useful for elucidating transmission pathways while sequencing entire viral genomes provides additional informative sites and often results in better-resolved phylogenies. We explore how whole-genome sequencing can benefit investigations of AMDV transmission by reconstructing the relationships between AMDV field samples from a Danish outbreak. We show that whole-genome phylogenies are much better resolved than those based on the partial NS1 gene sequences extracted from the same alignment. Well-resolved phylogenies contain more information about the underlying transmission trees and are useful for understanding the spread of a pathogen. In the main case investigated here, the transmission path suggested by the tree structure was supported by epidemiological data. The use of molecular clock models further improved tree resolution and provided time estimates for the viral ancestors consistent with the proposed direction of spread. It was however impossible to infer transmission pathways from the partial NS1 gene tree, since all samples from the case farms branched out from a single internal node. A sliding window analysis showed that there were no shorter genomic regions providing the same phylogenetic resolution as the entire genome. Altogether, these results suggest that phylogenetic analyses based on whole-genome sequencing taking into account sampling dates and epidemiological data is a promising set of tools for clarifying AMDV transmission.

INTRODUCTION

Aleutian mink disease (AMD), also referred to as plasmacytosis, is the most important disease in the mink farming industry worldwide. The disease affects mink of all ages and is caused by Aleutian mink disease virus (AMDV), a single-stranded DNA virus belonging to the family *Parvoviridae* [1]. Like other parvoviruses, the AMDV genome consists of two large ORFs and two smaller ones, which by alternative splicing encode three non-structural (NS1, 2 and 3) and two structural viral proteins (VP1 and 2) [2, 3]. Infection results in a harmful activation of the immune system, leading to hypergammaglobulinaemia and systemic vascular diseases such as glomerulonephritis. Animal welfare is reduced and infected animals either die due to organ failure or become persistently infected carriers, transmitting the virus within and between herds [4].

In Denmark, AMDV is monitored by a mandatory national control programme [5], which briefly requires all farms to conduct serology-based screening at regular intervals according to the region's disease status. Positive farms undergo more intensive monitoring and are encouraged to depopulate followed by thorough cleaning and disinfection of the farm. Given these regulations and due to the fact that parvoviruses are highly contagious and very resistant to environmental factors, AMDV protection and prevention imposes large costs on the mink farmers [4].

Received 20 June 2016; Accepted 15 March 2017
Author affiliations: [1]Kopenhagen Diagnostics, Kopenhagen Fur, Glostrup, Denmark; [2]Department of Bioinformatics, Technical University of Denmark, Lyngby, Denmark; [3]National Veterinary Institute, Technical University of Denmark, Frederiksberg, Denmark.
***Correspondence:** Emma E. Hagberg, eha@epista.com
Keywords: Aleutian mink disease virus (AMDV); whole-genome sequencing; next-generation sequencing (NGS); phylogeny; viral outbreak investigation.
Abbreviations: AMD, Aleutian mink disease; AMDV, Aleutian mink disease virus; ESS, effective sample size; GFAV, Gray fox amdoparvovirus; MCC, maximum clade credibility; MCMC, Markov-chain Monte Carlo; MRCA, most recent common ancestor; NGS, next-generation sequencing; NS, non-structural; VP, viral protein.
†Present address: Epista Life Science A/S, Hørsholm, Denmark.
Accession numbers: KU856560, KU856561, KU856562, KU856563, KU856564, KU856565, KU856566, KU856567, KU856568, KX404887, KX404888, KU856571, KU856572, KU856573, KX404889, KX404890, KX404891, KX404892, KX404893, KX404894, KX404895, KX404896, KX404897, KX404898, KX404899, KU856580, KU856569, KU856570, KX404900, KX404901, KX404902, KX404903, KX404904, KX404905, KX404906, KX404907, KX404908, KU856574, KU856575, KU856576, KX404909, KX404910, KX404911, KU856577, KX404912, KX404913, KX404914, KX404915, KX404916, KU856578.

Previous molecular studies of AMDV strains circulating in Denmark have primarily been based on Sanger sequencing of a part of the NS1 gene [6–9]. Such short and relatively conserved regions are useful for diagnostic purposes and for exploring more distant relationships between strains; however, due to the small number of informative sites, studies based on the partial NS1 gene have resulted in phylogenetic trees with low resolution with limited use for exploring outbreaks and discerning the transmission routes of AMDV *between* farms.

In this paper we demonstrate the strength and applicability of using whole-genome sequence for reconstructing phylogenetic relationships in a case of AMDV transmission among three Danish mink farms. The increased genetic information from the entire viral genomes may be used to investigate routes of viral transmission in more detail [10, 11]. Whole-genome data were obtained using next-generation sequencing (NGS), which has previously been a useful tool for this purpose [12, 13]. In connection with the phylogenetic analysis, we further explored the use of molecular clock models, which allowed us to estimate the age of the ancestors of the viruses from individual farms – thus generating information crucial for tracking the source of new outbreaks.

RESULTS

Sequence analysis

The data quality was overall high with approximately 99 % of the reads mapping to the AMDV-G reference genome (data not shown), and as previously observed, the homopolymeric region between nt 2470–2520 caused a dip in read coverage, but did not affect downstream analysis [3, 14]. The average read depth for each sample is shown in Table 1. The nucleotide diversity (average number of nucleotide differences per site) between the recent Danish isolates was relatively low: π=0.0062, SE=0.00032 for the partial NS1 gene and π=0.0043, SE=0.00014 for the whole-genome sequences. The individual mean pairwise differences between the sequences collected at each farm are reported in Table 1.

All full-length AMDV field strains were analysed for the presence of recombination using SimPlot [15] and all methods implemented in the RDP4 software package [16]. No recombination was detected between the sequences from the case farms (A, B and C), the remaining Danish isolates or between the Danish and international sequences. The SimPlot analysis reflected the overall low pairwise distances in the alignment and in line with previous studies indicated a slightly higher variability in the first ORF [3, 17].

Phylogenetic results

Evaluating the use of whole-genome sequences for reconstructing phylogenies

The two alignments, based on the partial NS1 gene and whole-genome sequences, were used as the input for the phylogenetic analyses. The best-fitting nucleotide substitution model for the partial-gene dataset was the so-called HKY model that distinguishes between transition and transversion rates as well as allowing for unequal base frequencies [18], while for the whole-genome dataset it was the HKY model with a proportion of invariant nucleotide sites and a gamma-rate distribution (HKY+IG). The phylogeny based on the partial-gene dataset showed all farm A, B and C sequences branching our from a single node (a polytomy) together with farm I; it was therefore unhelpful for estimating phylogenetic relationships and inferring transmission routes (see Fig. 1). The same conclusion applied to the remaining farms.

Analysis of the whole-genome dataset resulted in a better-resolved phylogeny with fewer polytomies and displayed high posterior clade probabilities (Fig. 1b). Sequences originating from farms B and C formed sub-trees within the tree of farm A. This is consistent with the hypothesis that farm A was infected first and that the infection then spread to farms B and C – an idea further supported by the clock-model analysis and the prevalence data (see below). The remaining Danish outbreak-derived sequences mainly clustered according to the farm from which they were sampled. The overall tree topology resembled that seen in previous studies [7, 19], with the Danish strains forming a clade of their own and the global strains all being placed as outgroups (Fig. 2). Our results illustrate that the whole-genome sequences contain additional important additional genetic information that improves the phylogenetic signal compared to the partial NS1 gene fragment used in Denmark. The considerably higher tree resolution obtained when using the whole-genome alignment enables us to begin understand these outbreaks in greater detail and investigate the route by which the infection spreads within a country.

Although the partial NS1 gene region seems to be insufficient for a robust phylogenetic analysis, it is possible that other short genomic regions might be useful for phylogenetic analysis. This could potentially be advantageous due to the somewhat easier workflow involved in Sanger sequencing and subsequent analysis of a single PCR fragment compared to the several steps involved in obtaining and analysing large collections of NGS reads. We therefore performed a sliding window analysis with the purpose of quantifying the phylogenetic information content in different subsections of the AMDV genome. Manual inspection of the whole genome alignment showed that nucleotide changes were located over the entire genome (suggesting that all the data are necessary to obtain full resolution), however with a slightly higher diversity in its first half. In order to stringently quantify the phylogenetic signal in different sub-sections of the alignment, 400 bp windows spaced at 25 bp intervals were extracted from the whole genome alignment (i.e. the windows started at positions 1, 25, 50, 75, 100, etc., with the first covering 1–400, the second 25–424, etc.). A window size of 400 was chosen as this corresponds to a typical PCR fragment size and could easily be Sanger-sequenced. For each of the resulting 179 sub-alignments, a tree was reconstructed using a full Bayesian phylogenetic analysis, and subsequently, the relative resolution for each window was measured in two different, but related, ways:

Table 1. Overview of sequences generated in this study

Sampling dates are indicated as yyyy-mm-dd, sequence lengths are either full length (4369 bp) or nearly full length (3198 bp), average read depth is the number of reads per nucleotide position, and the intra-farm diversity is presented as the mean pairwise sequence difference (pi) and its standard error (SE).

Farm	Mink no.	Sampling date (yyyy-mm-dd)	Length (bp)	Average read depth	GenBank acc. no.	Intra-farm diversity (mean pi, SE)
A	1	2014-11-21	4369	10959	KU856560	0.001068, 0.000257
	2	2014-11-21	4369	12912	KU856561	
	3	2014-11-21	4369	12627	KU856562	
B	1	2014-11-14	4369	12768	KU856563	0.00305, 0.000062
	2	2014-11-14	4369	6008	KU856564	
	3	2014-11-14	4369	14170	KU856565	
C	1	2014-11-14	4369	14211	KU856566	0.001373, 0.000285
	2	2014-11-14	4369	14238	KU856567	
	3	2014-11-14	4369	8015	KU856568	
D	1	2014-11-19	4369	3267	KX404887	0.000870,
	2	2014-11-19	4369	1001	KX404888	0.000137
	3	2014-02-10	4369	1496	KU856571	
	4	2014-02-10	4369	1339	KU856572	
	5	2014-02-10	4369	765	KU856573	
E	1	2014-05-08	4369	1283	KX404889	0.00458,
	2	2014-05-08	3198	1967	KX404890	0.000187
	3	2014-05-08	4369	1438	KX404891	
F	1	2014-11-11	3198	341	KX404892	0.00458,
	2	2014-11-11	3198	1040	KX404893	0.000187
	3	2014-11-11	4369	1174	KX404894	
G	1	2014-11-10	3198	1290	KX404895	0.000610,
	2	2014-11-10	3198	1277	KX404896	0.000125
	3	2014-11-10	4369	2263	KX404897	
H	1	2014-11-11	4369	2337	KX404898	0.001465,
	2	2014-11-11	3198	1155	KX404899	0.000220
	3	2014-11-11	3198	19014	KU856580	
	4	2014-11-21	4369	12655	KU856569	
	5	2014-11-21	4369	13965	KU856570	
I	1	2014-03-26	4369	2801	KX404800	0.000305,
	2	2014-03-26	4369	4338	KX404801	0.000125
	3	2014-03-26	4369	2030	KX404802	
J	1	2014-03-26	4369	2967	KX404803	0.000763,
	2	2014-03-26	4369	2546	KX404804	0.000125
	3	2014-03-26	4369	3257	KX404805	
K	1	2014-11-25	4369	2708	KX404806	0.001221,
	2	2014-11-25	4369	239	KX404807	0.000311
	3	2014-11-25	4369	338	KX404808	
L	1	2014-02-24	4369	23798	KU856574	0.000381,
	2	2014-02-24	4369	24082	KU856575	0.000080
	3	2014-02-24	4369	11615	KU856576	
	4	2014-02-24	4369	11903	KX404809	
	5	2014-02-24	4369	27274	KX404810	
	6	2014-02-24	4369	8565	KX404811	
M	1	2014-06-19	4369	2118	KU856577	0.000458,
	2	2014-06-19	4369	1341	KX404812	0.000072
	3	2014-09-19	4369	2203	KX404813	
	4	2014-09-19	4369	1974	KX404814	
	5	2014-09-19	4369	1587	KX404815	
N	1	2004-10-15	4369	8888	KX404816	–
O	1	2004-05-27	4369	10419	KU856578	–

(1) by counting the number of internal nodes in the tree – the better the resolution of a tree, the fewer polytomies and hence the more internal nodes the tree will have; and (2) by counting the number of fully resolved internal nodes (i.e. internal nodes having exactly two offspring) – a fully resolved tree will have two offspring for all internal nodes. These numbers were then compared to the corresponding measurements made on the whole-genome phylogeny (i.e. we computed the ratio between the values based on each individual sub-alignment and the value in the tree made from the full alignment) and the two measurements gave very similar results (data not shown). Fig. 3 shows the relative resolution measured using the

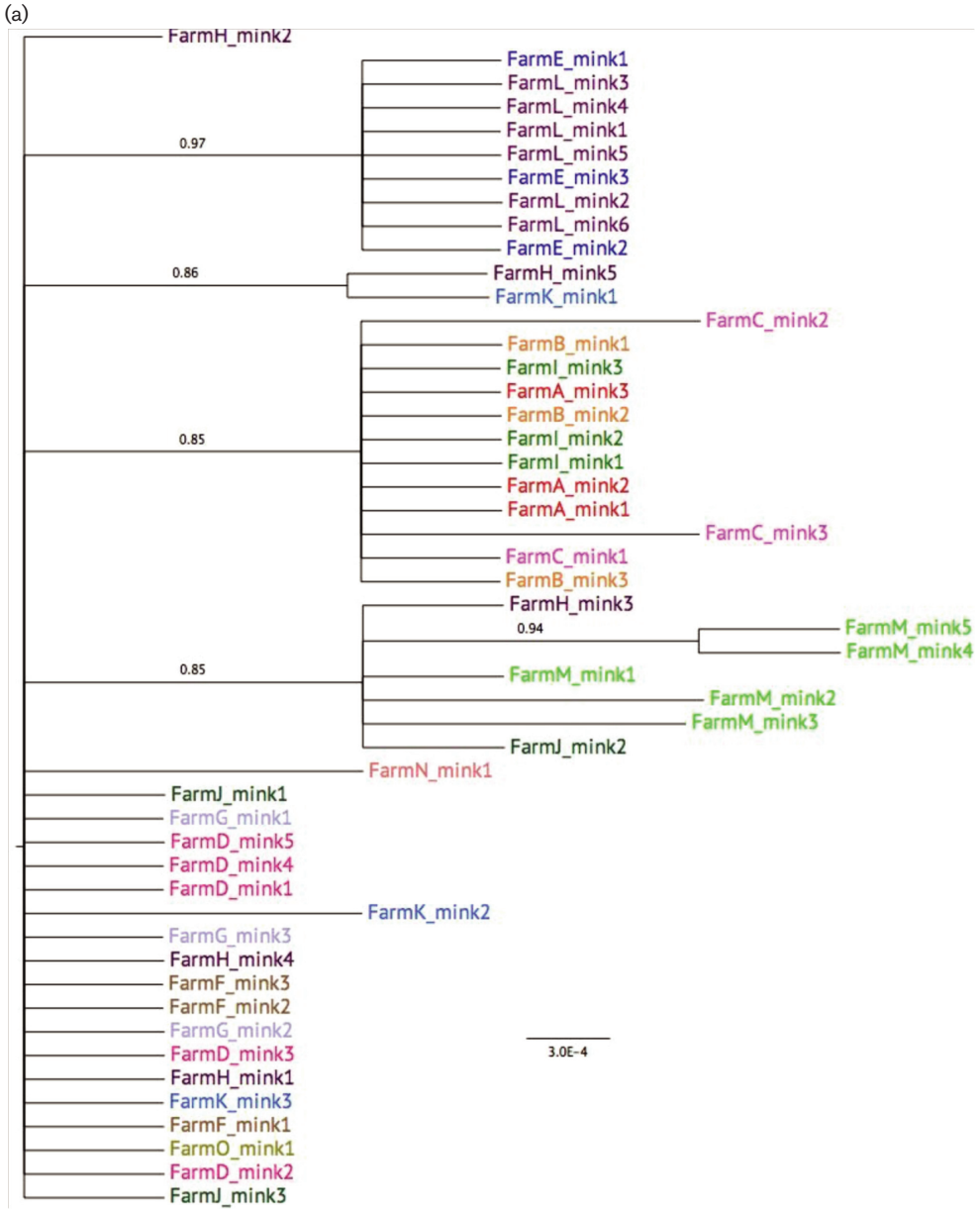
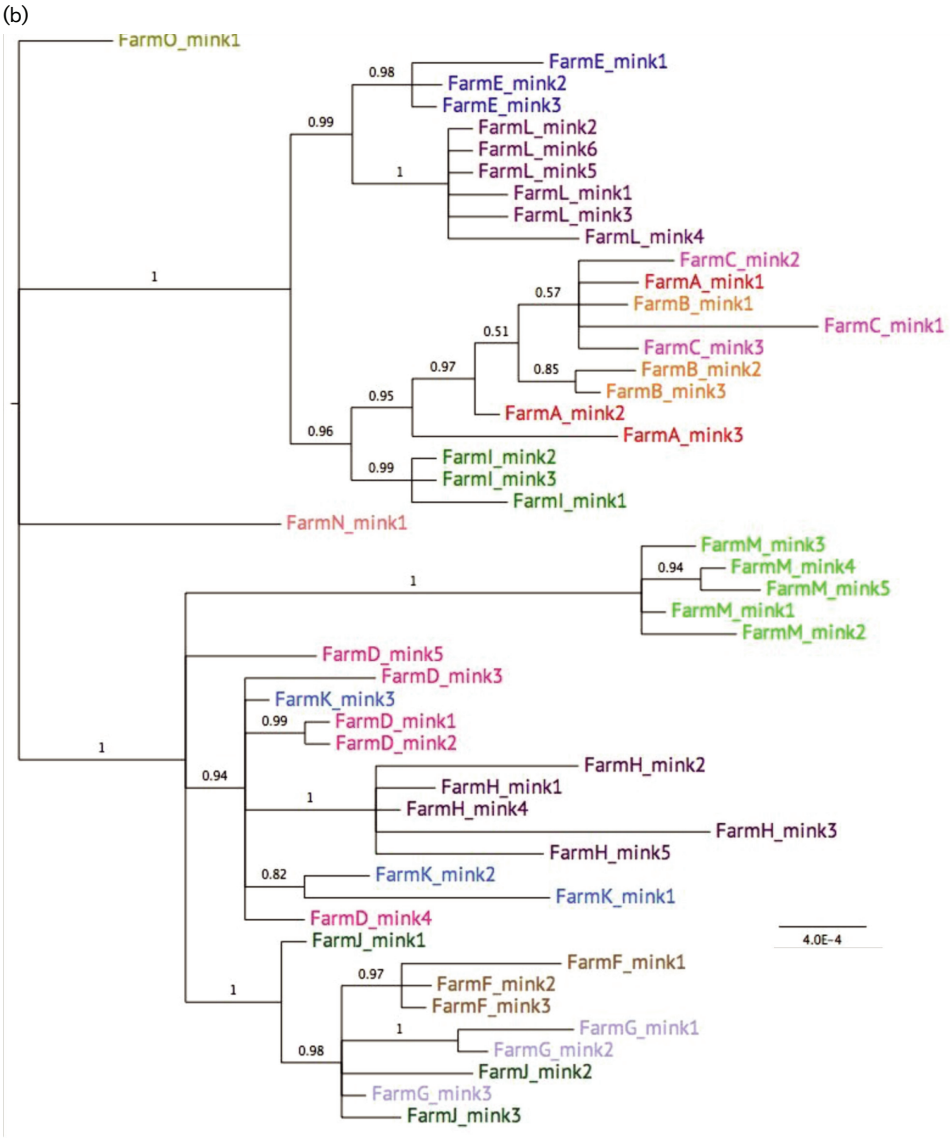


Fig. 1. Comparison between the partial NS1 gene and whole-genome phylogenies. Phylogenetic trees based on the partial NS1 gene dataset (a) and the whole-genome dataset (b) constructed in MrBayes version 3.2 applying the HKY model and estimating the number of invariable sites and gamma-rate distributions from the data. The Markov-chain Monte Carlo (MCMC) simulations were run for 50 million iterations. Branch labels represent posterior probabilities for each clade (Bayesian support values) and branch lengths represent substitutions per site, as indicated by the scale bars.

Fig. 1. (cont.)



number of internal nodes as a function of the start of the 400 bp window. Windows in the first half of the genome result in higher resolution, which is in agreement with the observed higher variability in this region. In particular, a region around approximately nt 1100–1700 (corresponding to the 3′ end of the NS1 gene) displayed good resolution. The segment typically used for partial NS1 gene sequencing is in this alignment located between nt 605–932, and corresponds to a dip in the curve of resolving power (Fig. 3). It should be noted that this gene segment was not originally selected for the purpose of maximizing phylogenetic signal content, but rather for reliable PCR amplification and diagnostic purposes, and thus our results emphasizes that the partial NS1 sequence is not well suited for investigating the spread of infection within Denmark. From our analysis, it is also clear that there is no single window of the genome that provides a resolution close to that of whole-genome sequencing, and even the best windows obtained at most 60 % of the resolution compared to using the whole-genome.

Estimating divergence times using sampling dates

The whole-genome dataset presented here was characterized by low diversity and a phylogeny with a shallow root and low levels of rate variation between its branches (mean nt difference=0.0033, SE=0.0001). Such datasets are often well described by a strict molecular clock and simple coalescent population growth, especially when the dataset represents a population subsample as in the present study [20, 21]. We used the BEAST2 software package to simultaneously estimate divergence times and absolute rates of molecular

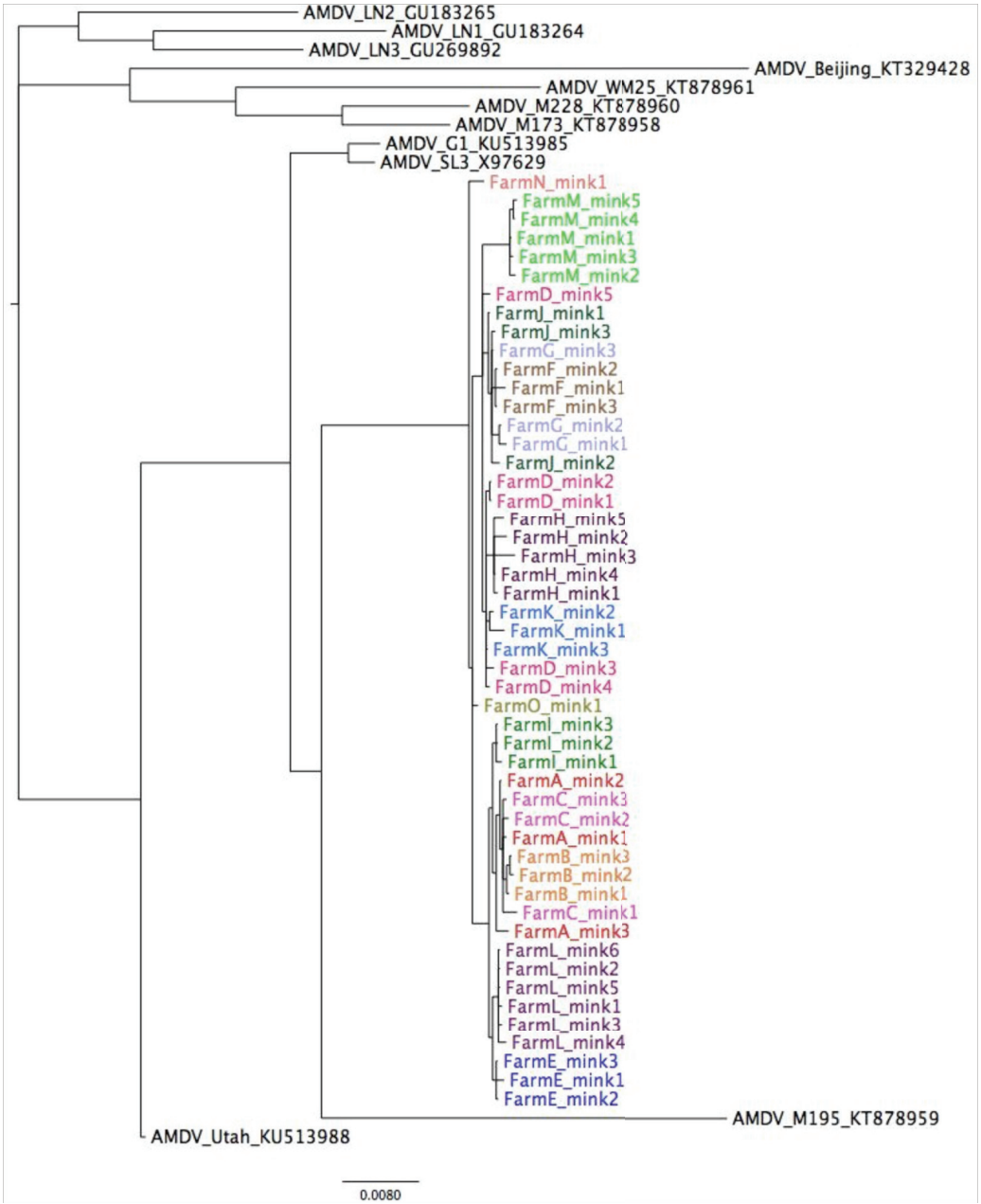


Fig. 2. Phylogenetic tree relating the whole-genome sequences generated in the present study in a global context. The tree was constructed using MrBayes applying the HKY model and estimating the number of invariable sites and gamma-rate distribution from the data and an MCMC run for 100 million iterations. Branch lengths represent substitutions per site as indicated by the scale bar. The tree was rooted on Gray Fox Amdoparvovirus (GFAV), which was removed from the summary tree to improve visibility.

evolution by including sampling dates in the phylogenetic analysis. BEAST2 works within a Bayesian paradigm and produce a probability distribution over possible values of the estimated times and rates.

The results of the time-stamped analysis strongly indicated that the most recent common ancestor (MRCA) for all sequences isolated from farm A was older than the MRCA for sequences isolated from farms B and C (Fig. 4). Specifically, the Bayesian analysis indicated with a posterior probability of 99.9 % that the farm A ancestor was the oldest. The estimated median age of the farm A MRCA was 1.4 years old (95 % credible interval=0.74–2.26 years), while the median age of the MRCA for farms B and C was 0.92 years (95 % credible interval=0.49–1.5 years). The sequences from farms B and C furthermore formed a sub-tree within the tree spread out by the farm A sequences. These observations were all consistent with the hypothesis that the viruses were

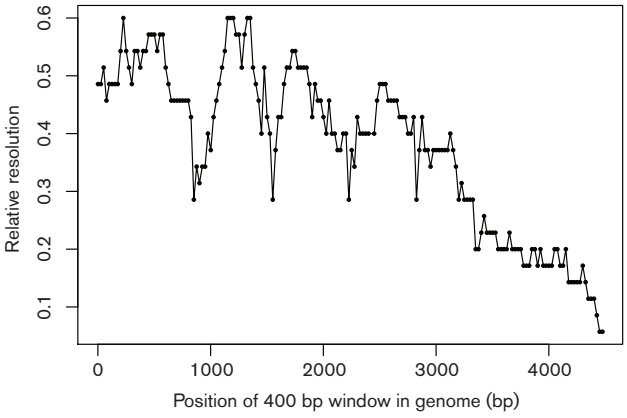


Fig. 3. Quantification of phylogenetic resolution across the AMDV genome. Bayesian phylogenetic trees were created using extractions of 400 bp partitions spaced at 25 bp intervals across the whole-genome alignment. The relative resolution was measured by comparing the number of fully resolved internal nodes in each partition to the full genome (*y*-axis), and plotted as a function of the starting position of the 400 bp window relative to the AMDV-G genome (x-axis).

transmitted from farm A to farms B and C, but not the opposite, and further supported the conclusions from the undated analyses above.

DISCUSSION

In this study, whole-genome sequencing and phylogenetic analyses were investigated and demonstrated to be valuable tools for determining AMDV transmission routes between mink farms. Previous phylogenetic analyses of AMDV in Denmark have been based on partial NS1 gene sequences. While a short and relatively conserved genomic region like this is suitable for diagnostic purposes, whole-genome sequences, which are longer and include more informative regions of the genome, are better suited for obtaining high phylogenetic resolution. We demonstrated this benefit by directly comparing whole-genome and partial NS1 gene phylogenies constructed from the very same data – a set of full-length AMDV sequences from which the sections corresponding to the previously used NS1 region were cut. The viral isolates were sampled during a small AMDV outbreak in three simultaneously infected farms (A, B, and C) in close geographical proximity to each other. A contemporary test population was generated by sampling an additional 12 simultaneously AMDV-infected Danish farms, and the data were put into context by including all at the moment available international full-length AMDV field isolates in the analysis. The serological test history showed that farm A had a much higher AMDV prevalence than its neighbours B and C, consistent with the idea that farm A became infected before farms B and C, and that the virus was transmitted from farm A to the other farms.

Based on the partial NS1 gene phylogeny, it was impossible to differentiate between transmission pathways since all sequences from farms A, B, and C formed a polytomy, meaning they branched from a single common internal node in the tree. Using the whole-genome sequences provided much higher phylogenetic resolution, and the sequences from both farms B and C consistently clustered within the farm A subtree, thus supporting the epidemiological transmission hypothesis that transmission was from farm A to farms B and C, and not the opposite. This direction of spread was further supported by the Bayesian clock model-based analysis of the relative dates of the MRCAs for the three case farms: the analysis indicated there was a high probability that farm A has the oldest viral ancestor, and hence was infected prior to the other farms. The use of whole-genome sequences also increased the phylogenetic resolution of the remaining farms regardless of substitution or tree model (data not shown), and we therefore concluded that whole-genome sequencing is a promising tool for identifying routes of AMDV transmission between farms.

It is important to realize that viral phylogenies and transmission trees are two different things, and that their structures can be quite dissimilar. This is because the physical transmission of a progeny virus must happen some time after it has genetically split from its parental lineage, and thus the internal nodes in the phylogeny (which correspond to the event at which two viral lineages split) will be further back in time than the equivalent internal nodes in the transmission tree (which correspond to transmission of a viral lineage to another animal). The branching order of the trees can also differ, since a single animal host contains many related viruses, and the order in which different lineages are transmitted to other animals is not necessarily the same order in which these lineages split on the viral phylogeny [22]. However, if samples are collected close enough in time to the transmission event, it should be possible to establish the direction of spread with more certainty [11, 23]. Furthermore, if sampling rates are low (i.e. if relatively few animals are sampled from each farm), then the timing of coalescent events in the phylogeny is very similar to the timing of transmission events and the structures of the two trees will be quite similar [24]. Under all circumstances, better resolution of the viral phylogeny will provide more information about the underlying transmission tree. In the present study, it should be taken into consideration that the farms in this enzootic area of Denmark have a history of culling their entire animal populations at the end of every production season and therefore a linear rate of viral evolution cannot be assumed. Despite the fact that adding sampling dates improved the resolution of the tree, it should be kept in mind that the inferred clock-rate, and hence ages of the MRCAs, does not necessarily reflect the real underlying mutation rate. The results presented here (based on samples from 2014) outlined this challenge as the MRCA of farms A, B, and C was estimated to be between 1.2 and 5.2 years, despite that the mink populations in these farms were culled at the end of the previous season (i.e. 2013). It cannot be ruled out that the virus infecting farm A was not a new virus, but originated from a previous outbreak and had persisted in the

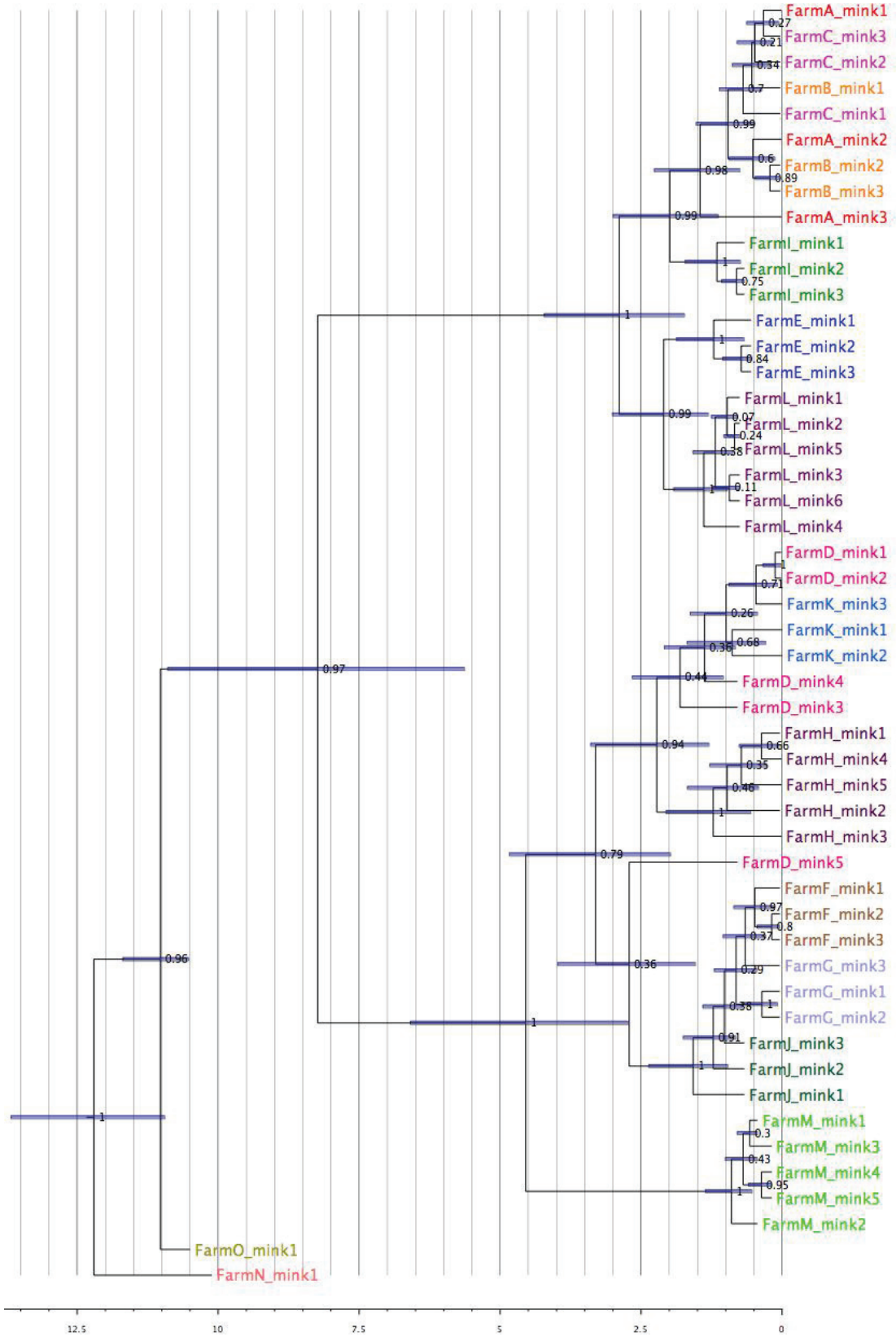


Fig. 4. Time-calibrated phylogenetic tree. The tree was constructed in BEAST2 using the HKY model and estimating the number of invariable sites and the gamma-rate distribution from the data, and applying a strict molecular clock and an exponential coalescent population

model. The MCMC chain was run for 100 million iterations. Node labels represent posterior probabilities (Bayesian support values) and node bars (blue lines) the 95 % confidence interval for the age distribution of each node. The *x*-axis represents time in years.

environment and later re-infected the new animals at the farm. Overall, the results from this study illustrate that phylogenetic analyses in relation to AMDV outbreaks will be most valuable in recently infected farms and that the time-stamped model is useful for inferring relative time points for the transmission events.

In addition to carefully choosing the parameters for the phylogenetic reconstruction process (e.g. substitution model and priors), the input data – that is the samples and sequences – are important. Sequencing of PCR-amplified DNA might not necessarily capture the full intra-individual viral variation due to clonal amplification. However, the use of a PCR pre-amplification step is common practice and due to e.g. abundant host DNA and low amounts of viral DNA, a feasible approach for the purpose of detecting and typing AMDV [3]. Possible within-farm variation could theoretically be addressed by sequencing additional samples per farm and by including a contemporary test population in the analysis, as was done in the present study. However, due to practical circumstances in regard to farm operations, it will rarely be possible to get more than one or two samples per farm, but the low intra-farm diversity presented in the present study suggests it is nevertheless possible to tease out the inter-farm spread by including additional farms and adding sampling dates in the analysis.

The practical operations of a mink farm, e.g. open barns allowing for plentiful ventilation, possible wildlife access, external feed-supply transportation routes and the proximity between the farms illustrated by the case farms A and B (Fig. 5), can impede biosafety and should be considered carefully in the daily routines at the farm. A study of avian influenza in livestock showed that the wind direction on the date of transmission was correlated to the between-farm viral spread [25]. However, as mentioned above, the specific time points for the transmission events could not be determined in this study and therefore the impact of external factors such as the wind will remain a speculation. Presumably, neither a constant or exponential tree population model, nor a strict molecular clock, is the best way to describe the dynamics in a viral population. But on the other hand, to successfully infer relevant parameters using, e.g., the more complex birth-death models, additional knowledge about population parameters and a larger dataset would be required. The number of publicly available whole-genome AMDV sequences is currently very limited, and it was just recently that Canuti *et al.* [19], in a study of field strains originating from Canada, China and Germany, demonstrated inconsistent phylogenies due to large variation between tree structures over the viral genome and the presence of potential recombination events. Thus, we suggest that future studies should aim to investigate larger test populations sampled from broader timespans and from different countries in order to map the entire AMDV genomic diversity at a global level, thereby generating data that would benefit all mink-farming countries regardless of prior genomic surveillance strategy.

In conclusion, this study illustrates that whole-genome sequencing is better suited for reconstructing high-resolution phylogenetic relationships between AMDV isolates compared to shorter gene fragments such as the partial NS1 gene fragment currently used for AMDV typing in Denmark. Furthermore, by confirming an epidemiological transmission route hypothesis between three case farms, we show that whole-genome phylogenies supplement epidemiological data, such as AMDV prevalence and test history of the farms, to indicate the direction of transmission, thus suggesting a framework that could become an important tool to identify inter-farm spread of AMDV.

METHODS

Farms A, B and C

We investigated a case of AMDV transmission in a small Danish AMDV outbreak where there was a strong *a priori* hypothesis about the route of viral transmission based on epidemiological data (Kopenhagen Fur, personal communication). Farms A, B and C all tested serologically positive for AMDV in November 2014. In addition, farm A had tested positive in August 2014 with an estimated prevalence of 21.4 %, in September the prevalence of farm C was estimated to 0.4 %, while farm B tested negative for AMDV in 2014 (overview in Fig. 5). These prevalence estimates were based on serological testing of a percentage of the breeding animals at certain intervals according to the Danish control programme [5]. The much higher AMDV prevalence on farm A compared to farms B and C at the time of testing in November 2014 formed the basis of the initial hypothesis that farm A transmitted AMDV to farms B and C.

To put these case farms into context, a contemporary test population consisting of mink sampled from other Danish farms tested positive for AMDV in 2014, was included in the analysis. Potential within-farm variation was addressed by, when possible, sequencing more than one sample per farm. However, due to practical circumstances in regards to farm operations, it is rarely possible to access more than one or two samples per farm, but this number should be sufficient to estimate between-farm variation if the dataset is large enough. The geographic locations of the farms and their AMDV prevalence at their most recent AMD serology test is shown in Fig. 5. Sequences from farms N and O, both sampled in 2004, were included based on the hypothesis of being sufficiently evolutionarily distant to root a time-stamped tree, but not being so different that they would introduce

	Farm A	Farm B	Farm C	Farm D	Farm E	Farm F	Farm G	Farm H	Farm I	Farm J	Farm K	Farm L	Farm M	Farm N	Farm O
Breeding animals 2014	1600	800	2700	2490	1150	3900	2000	5800	1300	200	2500	1000	2450	600	2200
AMDV prevalence (%)	21,4	1,0	0,0	6,9	0,0	0,1	33,9	97,1	0,0	0,0	19,1	0,0	0,5	n.a.	n.a.

Fig. 5. Map of Denmark and overview of the included farms. Insert shows close-up of farms A, B and C. For each farm the number of breeding animals during the season of 2014 and the AMDV prevalence as a percentage of positive animals in the most recent antibody test are indicated. Map created using Tableau Desktop version 9.2 and insert from Google DigitalGlobe 2015.

Table 2. Overview of sequences retrieved from NCBI GenBank

Viral strain	Country	Length (bp)	GenBank accession no.	Reference
AMDV-M173	Canada	4200	KT878958	[19]
AMDV-M195	Canada	4161	KT878959	[19]
AMDV-M228	Canada	4164	KT878960	[19]
AMDV-WM25	Canada	4169	KT878961	[19]
AMDV-Beijing	China	4802	KT329428	[33]
AMDV-LN1	China	4543	GU183264	[34]
AMDV-LN2	China	4566	GU183265	[33]
AMDV-LN3	China	4566	GU269892	[33]
AMDV-SL3	Germany	4718	X97629	[35]
AMDV-Utah	Antigen	4369	KU513988	[3]
AMDV-G#1	Antigen	4369	KU513985	[3]
GFAV	USA	4441	JN202450	[36]

long branch lengths that could affect estimation tasks in Bayesian time-line analyses [26].

Viral samples and preparation

One blood sample per animal was collected from a total of 48 animals originating from farms A–M, in addition to one archive blood sample from one animal from farms N and O, respectively (Table 1). Total DNA was extracted from each blood sample using QIAmp MinElute Virus Spin Kit (Qiagen, Hildren, Germany) and eluted in 50 uL low TE-buffer according to the manufacturer's instructions. The viral DNA was PCR-amplified as described previously [3] and submitted to the Technical University of Denmark (DTU) Multi-Assay Core (Lyngby, Denmark) for library preparation and sequencing on a 318 chip using the Ion Torrent PGM (Life Technologies, Carlsbad, CA) according to the manufacturer's instructions. Raw data in fastq format were processed as described previously [3]. Briefly, the steps prior to sequence assembly included QC, trimming, error correction and mapping to the AMDV-G reference with accession no. NC_001662, followed by naming each sequence according to Table 1. For a global comparison, previously published complete coding sequences for AMDV field strains originating from Canada, China and Germany, the complete coding sequences for the antigen strains AMDV-Utah and AMDV-G, in addition to the closely related Gray fox amdoparvovirus (GFAV), were retrieved from NCBI GenBank (overview and references in Table 2).

Sequence analysis

Intra-farm diversity was estimated by calculating the mean pairwise nucleotide distance and its corresponding standard error (SE) between the sequences collected from each farm [27]. The full-length field strains were analysed using SimPlot [15] and all methods implemented in the RDP4 software package [16], i.e. RDP, GENECONV, BootScan, MaxChi, Chimera, SiScan, 3Seq and LARD, to investigate recombination and to describe variability across the genome.

Phylogenetic reconstructions

Evaluating the use of whole-genome sequences for reconstructing phylogenies

The whole-genome sequences, including GFAV for rooting, were aligned at nucleotide level using MAFFT V.7.205 [28] and converted to nexus format. The 327 bp region corresponding to the DNA sequence flanked by the primers used for 'conventional PCR' amplification [8] was extracted from the alignment, thereby creating two datasets: 'partial NS1 gene' and 'whole genome'. The best fitting substitution model for each dataset was selected using the program jModeltest [29]. The phylogenetic relationships were inferred in a Bayesian framework with Markov-chain Monte Carlo (MCMC) sampling in MrBayes version 3.2.3 [30], applying an HKY model with an estimated proportion of invariable sites, four gamma distribution rate categories and Dirichlet priors. The MCMC was run for 50 million generations for each of the datasets. The first 25 % of the samples was discarded as burn-in, and effective sample size (ESS) values above 400 for all parameters and standard deviation of split frequencies below 0.001 were considered as indications that the MCMC had converged successfully. FigTree version 1.4.2 was used for tree manipulations such as rooting and for visualization. To facilitate visualization, the outgroup sequence (GFAV) was removed from the maximum clade credibility (MCC) trees.

Estimating divergence times using sampling dates

The feasibility of using viral sampling dates to perform divergence time dating of the ancestors of the isolates involved in the outbreak was investigated in a Bayesian framework implemented in BEAST version 2.4.1 [31]. Often a simple model such as a strict molecular clock and a coalescent constant population growth prior size model is a useful starting point for analysis [20], especially if the dataset represents a subsample of the population as in the present study. In addition, both a strict and relaxed molecular clock, a coalescent constant and an exponentially growing tree population model was tested, and sampling was done with MCMC simulations run for 100 million generations to obtain estimates of the posterior distributions. The whole-genome dataset, excluding the GFAV isolate in order to avoid introducing long branch-lengths, was aligned as described above and used for this analysis. The MCMC log files were inspected for chain convergence based on the magnitude of ESS-values, and shapes of the traces and marginal posterior probabilities using Tracer version 1.6 [32]. Treeannotator version 1.8.2 was used for summarizing the tree log files, and FigTree version 1.4.2 (both distributed with the BEAST package [31] were used for tree manipulations such as rooting and visualization. The model was also run without data to confirm that the priors did not influence the results unduly.

Funding information
This work was funded by the Research Foundation of the Danish Fur Breeder's Association and the 'Innovation Fund Denmark' (award number: 16 479).

Acknowledgements
The co-workers at Kopenhagen Diagnostics, Kopenhagen Fur, are thanked for collecting the samples and for providing valuable background information about the farms' history.

Conflicts of interest
The authors declare there are no conflicts of interest.

Ethical statement
The work presented in the paper does not consist of experiments performed on humans or animals, but were performed using tissue samples.

References

1. **Bloom ME, Race RE, Wolfinbarger JB.** Characterization of Aleutian disease virus as a parvovirus. *J Virol* 1980;35:836–843.
2. **Bloom ME, Alexandersen S, Perryman S, Lechner D, Wolfinbarger JB.** Nucleotide sequence and genomic organization of Aleutian mink disease parvovirus (ADV): sequence comparisons between a nonpathogenic and a pathogenic strain of ADV. *J Virol* 1988;62:2903–2915.
3. **Hagberg EE, Krarup A, Fahnøe U, Larsen LE, Dam-Tuxen R *et al.*** A fast and robust method for whole genome sequencing of the Aleutian mink disease virus (AMDV) genome. *J Virol Methods* 2016;234:43–51.
4. **Decaro N, Buonavoglia C, Ryser-Degiorgis M-P, Gortázar C.** Parvovirus infections. In: Gavier-Widén D, Duff JP and Meredith A (editors). *Infectious Diseases of Wild Mammals and Birds in Europe*, 1st ed. Oxford, UK: Wiley-Blackwell; 2012. pp. 181–285.
5. Danish executive order 1447 of 15/12/2009. 2009. www.retsinformation.dk/Forms/R0710.aspx?id=129366.
6. **Christensen LS, Gram-Hansen L, Chriél M, Jensen TH.** Diversity and stability of Aleutian mink disease virus during bottleneck transitions resulting from eradication in domestic mink in Denmark. *Vet Microbiol* 2011;149:64–71.
7. **Leimann A, Knuuttila A, Maran T, Vapalahti O, Saarma U.** Molecular epidemiology of Aleutian mink disease virus (AMDV) in Estonia, and a global phylogeny of AMDV. *Virus Res* 2015;199:56–61.
8. **Jensen TH, Christensen LS, Chriél M, Uttenthal A, Hammer AS.** Implementation and validation of a sensitive PCR detection method in the eradication campaign against Aleutian mink disease virus. *J Virol Methods* 2011;171:81–85.
9. **Jensen TH, Hammer AS, Chriél M.** Monitoring chronic infection with a field strain of Aleutian mink disease virus. *Vet Microbiol* 2014;168:420–427.
10. **Escobar-Gutiérrez A, Vazquez-Pichardo M, Cruz-Rivera M, Rivera-Osorio P, Carpio-Pedroza JC *et al.*** Identification of hepatitis C virus transmission using a next-generation sequencing approach. *J Clin Microbiol* 2012;50:1461–1463.
11. **Metzker ML, Mindell DP, Liu XM, Ptak RG, Gibbs RA *et al.*** Molecular evidence of HIV-1 transmission in a criminal case. *Proc Natl Acad Sci USA* 2002;99:14292–14297.
12. **Kvisgaard LK, Hjulsager CK, Fahnøe U, Breum Solvej Ø., Ait-Ali T *et al.*** A fast and robust method for full genome sequencing of Porcine Reproductive and Respiratory Syndrome Virus (PRRSV) Type 1 and Type 2. *J Virol Methods* 2013;193:697–705.
13. **Jakhesara SJ, Bhatt VD, Patel NV, Prajapati KS, Joshi CG.** Isolation and characterization of H9N2 influenza virus isolates from poultry respiratory disease outbreak. *Springerplus* 2014;3:196.
14. **Quail MA, Smith M, Coupland P, Otto TD, Harris SR *et al.*** A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* 2012;13:341.
15. **Lole KS, Bollinger RC, Paranjape RS, Gadkari D, Kulkarni SS *et al.*** Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. *J Virol* 1999;73:152–160.
16. **Martin DP, Murrell B, Golden M, Khoosal A, Muhire B.** RDP4: detection and analysis of recombination patterns in virus genomes. *Virus Evol* 2015;1:1–5.
17. **Gottschalck E, Alexandersen S, Storgaard T, Bloom ME, Aasted B.** Sequence comparison of the non-structural genes of four different types of Aleutian mink disease parvovirus indicates an unusual degree of variability. *Arch Virol* 1994;138:213–231.
18. **Hasegawa M, Kishino H, Yano T.** Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* 1985;22:160–174.
19. **Canuti M, O'Leary KE, Hunter BD, Spearman G, Ojkic D *et al.*** Driving forces behind the evolution of the Aleutian mink disease parvovirus in the context of intensive farming. *Virus Evol* 2016;2:vew004.
20. **Drummond AJ, Syw H, Rawlence N, Rambaut A.** A Rough Guide to BEAST 1. 4. 2007:1–41.
21. **Brown RP, Yang Z.** Rate variation and estimation of divergence times using strict and relaxed clocks. *BMC Evol Biol* 2011;11:271.
22. **du Plessis L, Stadler T.** Getting to the root of epidemic spread with phylodynamic analysis of genomic data. *Trends Microbiol* 2015;23:383–386.
23. **Bowman BH, White TJ.** Molecular epidemiology of AIDS. In: Schochetman G and George JR (editors). *AIDS Testing*. New York, NY: Springer New York; 1994.
24. **Ypma RJ, van Ballegooijen WM, Wallinga J.** Relating phylogenetic trees to transmission trees of infectious disease outbreaks. *Genetics* 2013;195:1055–1062.
25. **Ypma RJ, Jonges M, Bataille A, Stegeman A, Koch G *et al.*** Genetic data provide evidence for wind-mediated transmission of highly pathogenic avian influenza. *J Infect Dis* 2013;207:730–735.
26. **Drummond AJ, Bouckaert RR.** *Bayesian Evolutionary Analysis with BEAST 2*. Cambridge University Press; 2015. p. 260.
27. **Pedersen AG.** *seqlib.py*. 2012.
28. **Katoh K, Standley DM.** MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 2013;30:772–780.
29. **Darriba D, Taboada GL, Doallo R, Posada D.** jModelTest 2: more models, new heuristics and parallel computing. *Nat Methods* 2012;9:772.
30. **Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A *et al.*** MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol* 2012;61:539–542.
31. **Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu CH *et al.*** BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput Biol* 2014;10:e1003537.
32. **Rambaut A, Suchard M, Xie D, Drummond A.** 2014. Tracer v1.6. http://beast.bio.ed.ac.uk/Tracer
33. **Xi J, Wang J, Yu Y, Zhang X, Mao Y *et al.*** Genetic characterization of the complete genome of an Aleutian mink disease virus isolated in north China. *Virus Genes* 2016;52:463–473.
34. **Li Y, Huang J, Jia Y, du Y, Jiang P *et al.*** Genetic characterization of Aleutian mink disease viruses isolated in China. *Virus Genes* 2012;45:24–30.
35. **Schuierer S, Bloom ME, Kaaden OR, Truyen U.** Sequence analysis of the lymphotropic Aleutian disease parvovirus ADV-SL3. *Arch Virol* 1997;142:157–166.
36. **Li L, Pesavento PA, Woods L, Clifford DL, Luff J *et al.*** Novel amdovirus in gray foxes. *Emerg Infect Dis* 2011;17:1876–1878.

4.3. MANUSCRIPT 3

Genetic analysis of the entire genome of Aleutian Mink Disease Virus determines its evolutionary rate and confirms bottleneck due to control program.

Status: draft in preparation.

(page numbers are relative to paper)

Genetic analysis of the entire genome of Aleutian Mink Disease Virus determines its evolutionary rate and confirms bottleneck due to control program

Emma E. Hagberg[a,b,1*], Lars E. Larsen[c], Anders Krarup[a], Anders G. Pedersen[b]

[a]Kopenhagen Fur, Glostrup, Denmark
[b]Department of Bioinformatics, Technical University of Denmark, Lyngby, Denmark
[c]National Veterinary Institute, Technical University of Denmark, Frederiksberg, Denmark

*Corresponding author: Emma E. Hagberg
E-mail: eha@epista.com
[1] Current address: Epista Life Science, Slotsmarken 17, DK-2750 Hørsholm

Abstract word count:

Text word count:

Accession numbers: KY996892-KY997057

Abstract

Aleutian Mink Disease Virus (AMDV) is worldwide the most important virus in mink. It affects mink of all ages, impacts animal welfare, and imposes a burden to the farmers and the economy. In Denmark, AMDV is monitored through a national control program encouraging infected farms to depopulate. Genomic-based outbreak investigations have been hampered by low phylogenetic resolution, the lack of robust evolutionary rate-estimates, and a lack of consensus regarding which genomic region to sequence for molecular-based surveillance of AMDV.

In the present study 166 full-length AMDV genomes were next generation sequenced and their phylogeny inferred in a Bayesian phylogenetic framework to reconstruct their phylogenetic relationships and to describe the dynamics of AMDV in Denmark during the past decades. The strains clustered into two distinct populations, indicating a long time since divergence and possibly also a large number of unsampled hosts. The oldest cluster, the previously described Saeby strain, was subject to less selective pressure and lower genomic variation compared to the strains related to the recent

Danish AMDV outbreaks. A Bayesian skyline analysis of the Saeby strain revealed a sharp decrease in AMDV effective population size just a few years *after* the implementation of the Danish control programme in 1999.

Robust evolutionary rates for AMDV were estimated based on full-length sequences, and sites possibly linked to pathogenicity were identified. Supplemented with denser sampling, these tools could play an important role for determining the timing and origin of future AMDV outbreaks.

Keywords

Aleutian Mink Disease Virus (AMDV); next generation sequencing (NGS); whole genome sequencing; phylogeny; phylodynamics; molecular clock; selection pressure

1. Introduction

Aleutian Mink Disease (AMD), also referred to as Plasmacytosis, is worldwide the most important disease in the mink farming industry. The disease affects mink of all ages and is caused by Aleutian Mink Disease Virus (AMDV), a single-stranded DNA virus belonging to the family *Parvoviridae* (Bloom et al. 1980). Similar to other parvoviruses the AMDV genome consists of two large open reading frames (ORFs), two smaller ORFs, which by alternative splicing encode three non-structural (NS1, 2, and 3) and two structural viral proteins (VP1, and 2) (Bloom et al. 1988; Hagberg et al. 2016). Infection results in a harmful activation of the immune system leading to hypergammaglobulinaemia and systemic vascular diseases like glomerulonephritis. Animal welfare is reduced and infected animals either die due to organ failure or become persistently infected carriers capable of shedding the virus to its surroundings (Decaro, Nicola et al. 2012).

In Denmark AMDV is monitored by a national control program (Anon 2009), which briefly requires all farms to conduct serology-based screening at regular intervals according to the disease status of the region. Positive farms undergo more intensive monitoring and are encouraged to depopulate followed by thorough cleaning and disinfection of the farm. Given these regulations and the fact that parvoviruses are

highly contagious and resistant to environmental factors, AMDV protection and prevention imposes large costs to the fur industry.

Historically there has been no consensus regarding which part of the AMDV genome the various fur producing countries has used for genomic surveillance, and in most countries there is no systematic monitoring at all, only ad hoc sequencing of selected cases. A few larger studies of strains from Finland and Estonia have shed light on the molecular evolution of the VP2-gene (Leimann et al. 2015; Nituch et al. 2012; Wang et al. 2014), while most sequencing in Denmark traditionally has been limited to the partial-NS1 gene flanked by the same PCR-primers as used for diagnostic PCR (Jensen et al. 2011). In the season of 2015/2016, two new AMDV genotypes were detected on Danish farms, however, it was not possible to elucidate transmission routes (Ryt-Hansen et al. 2017) based on sequencing of the partial NS1 gene, suggesting that more thorough analysis of the entire AMDV genome is needed.

Next generation sequencing (NGS) of entire genomes followed by phylogenetic analyses is useful for confirming known patterns of viral spread and for revealing potential links suggested by epidemiological data (Valdazo-González et al. 2012; Metzker et al. 2002; Stadler et al. 2014). However, it was not until recently the entire genomes of AMDV field strains were sequenced using NGS (Hagberg et al. 2016), and the benefits of using these data in regards to elucidating transmission patterns between mink farms was demonstrated (Hagberg et al. 2017). The latter study was performed on contemporaneously sampled strains originating from a single AMDV cluster: the so-called “Saeby” strain, which has been circulating in Denmark for a long time (Christensen et al. 2011), and emphasized the need to investigate a larger set of samples.

Robust estimates for the molecular clock-rate of a virus are useful for characterising its evolution and for accurately tracing back in time the emergence of an outbreak and determining the age of the most recent common ancestor (MRCA) for a set of samples. Thus, the aim of this study was to apply phylogenetic and phylodynamic methods on whole genome sequences from AMDV field strains in order to i) estimate the molecular clock rate of AMDV, and ii) investigate the impact of the Danish control and depopulation program on AMDV population diversity and adapatation

over time.

2. Material and methods

2.1. Samples and sample preparation

Spleen or blood samples were gathered from known AMDV-infected animals originating from farms in Denmark, Poland, The Netherlands, or retrieved from the freezer archives in Kopenhagen Fur. From each sample total DNA was extracted and PCR amplified in a single fragment with the primers and protocols as described previously (Hagberg et al. 2016). Library preparation (400bp) and sequencing on a 318-chip using the Ion Torrent PGM™ (Life Technologies, Carlsbad, CA) was performed according to the manufactures instructions at the Technical University of Denmark (DTU) Multi-Assay Core (Lyngby, Denmark). A total of 166 samples were generated in this study (overview in Supplementary A), and an additional recently published full genome AMDV sequences were retrieved from GenBank: 42 from Denmark (Hagberg et al. 2017), three from China (Li et al. 2012), and four from Canada (Canuti et al. 2016) (overview in Supplementary A).

2.2. Sequence analysis

Briefly, raw-data in fastq-format were processed as follows: each sequence underwent QC, trimming and error correction prior to mapping to the AMDV-G reference with accession number NC_001662 as described previously (Hagberg et al. 2016) excluding the error-correction step, and the resulting consensus sequence for each viral sample was named according to Supplementary A. The consensus sequences, including those retrieved from GenBank, were aligned on nucleotide level using MAFFT (Katoh & Standley 2013) and converted to nexus format. The total alignment consisted of 215 samples. The best fitting substitution models for each data-set was determined using the bModelTest (Bouckaert 2015) implemented in Beast2.

2.3. Recombination analysis

Possible recombination events were searched for using Bootscan, Chimaera, GENOMCOV, maximum X^2, and SiScan methods implemented in the RDP4 software package using default parameters (Martin et al. 2015).

2.4. Temporal analysis

TempEst (Rambaut et al. 2016) was used to perform a root-to-tip analysis to test whether the samples exhibited adequately clock-like behaviour. TempEst does this by plotting the regression of the root-to-tip genetic distance inferred from maximum likelihood (ML) trees and the sampling dates. Linear relationship and small residuals indicates evolution can be described with a strict clock, while larger residuals suggest a relaxed molecular clock. Non-linear trends suggest that evolutionary rates have changed through time, and no trend implies there is no or little temporal signal and that data is unsuitable for molecular clock models. The ML trees were created using the ML plugin in Genious v.7.1. (Kearse et al. 2012) resampling with 100 bootstrap replications.

2.5. Phylogenetic analyses

2.5.1. Viral population diversity and evolution through time

The phylogenetic relationships were inferred in the Bayesian framework implemented in Beast v2.4.1 (Bouckaert et al. 2014) with Markov-chain Monte Carlo (MCMC) sampling run for 50 million generations to obtain estimates of the posterior distributions. The Hasegawa-Kishino-Yano (HKY) DNA model for nucleotide evolution with an estimated proportion of invariable sites and four gamma distribution rate categories was applied, and the following tree population models were tested: (i) a coalescent constant, (ii) a coalescent exponential, and (iii) a coalescent Bayesian skyline model. Following molecular clock models were tested: (i) a strict molecular clock, (ii) a relaxed molecular clock with an uncorrelated log-normal rate distribution, and (iii) a relaxed molecular clock with an uncorrelated exponential rate distribution. The trees were calibrated on sampling years, as this approach allowed for the previously published global sequences retrieved from GenBank to be included in the analysis.

The first 20% of the samples were discarded as burn-in. The MCMC log files were inspected for chain convergence, the magnitude of effective sample size (ESS) values, and shapes of the marginal posterior probabilities using Tracer v1.6 (Rambaut et al. 2014). ESS values above 400 for all parameters and even mixing in all chains were considered indications that the MCMC had converged successfully. The robustness of

the parameters was assessed by testing different combinations of molecular clocks, and demographic and phylogeographic diffusion models. In addition, different compositions of the datasets (i.e. the entire dataset or a single sequence per farm) were tested in order to confirm the estimates were not a product of sampling (Kühnert et al. 2011). Treeannotator v1.8.2 was used for summarizing the tree log files from each run, and FigTree v.1.4.2 (both distributed with the Beast-package (Bouckaert et al. 2014)) were used for tree manipulations such as rooting and visualisation.

2.5.2. Sliding window analysis

A sliding-window approach was undertaken, as described previously (Hagberg et al. 2017), to investigate the phylogenetic resolution in smaller sub sequences. Briefly, phylogenies were reconstructed using 400bp partitions distributed with 25bp increments along the complete alignment (n=215). The relative resolution for each partition was obtained by dividing the number of resolved nodes compared to a reference phylogeny based on the whole genome.

2.6. Selection pressure

The *overall ratio* of synonymous and non-synonymous substitutions (dN/dS) was calculated separately for NS1 and VP2 gene alignments using the Single Likelihood Ancestor Counting (SLAC) method available via the web interface Datamonkey (www.datamonkey.com) (Pond, S. L., Frost 2005). The selective pressure acting at the *individual codons* was assessed separately for the NS1 and VP2 genes for each cluster (A and BC), using four different likelihood based methods available at Datamonkey: the Single Likelihood Ancestor Counting (SLAC), the Fixed Effect Likelihood (FEL), the Internal FEL (IFEL), and the Random Effects Likelihood (REL). Since these methods perform somewhat differently, e.g. SLAC is known to be less sensitive, FEL is known for overestimating the number of sites, while MEME can predict sites under episodic diversifying selection, it is common practice to accept a codon as being under selection when predicted by at least three of the methods (Canuti et al. 2016). In cluster BC, 45 and 24 sequences were removed from the NS1 and VP2 alignments, respectively, due to insertions and deletions (indels) causing shifts in the major open reading frames. These indels were most likely artefacts as the consensus sequences represent an “average viral sequence” from each sample they

205 came from, and thus not necessarily a viable viral strain. Cluster A was reduced to
206 contain one sequence per farm per year because of the 75 sequence limit at
207 Datamonkey (Pond, S. L., Frost 2005).
208

209 # 3. Results

210 The final alignment (4425nt, corresponding to 92% of the AMDV genome) consisted
211 of 215 AMDV sequences originating from known AMDV-infected mink from farms
212 in Denmark, Poland, The Netherlands, and Canada, in addition to two wild mink from
213 Bornholm and one from Canada (Supplementary A), all sampled between 2004-201.
214 The 166 sequences generated in the present study are available in GenBank under the
215 accession numbers KY996892-KY997057.
216

217 ## 3.1. Recombination and temporal structure

218 Analysis using a range of methods did not indicate the presence of recombination in
219 the investigated viral sequences. To examine if the datasets exhibited adequate
220 temporal structure for estimation of clock-rates, root-to-tip analyses were performed,
221 and the resulting plots of the genetic distance versus isolation time-point (fig. 1) were
222 manually inspected. This analysis indicated that a clock model was not strongly
223 supported, and a cause for this, might be that the sequences were divided into two
224 distinct clusters by a very deep split, between the Saeby-cluster and the other cluster
225 containing the sequences isolated from the 2015/2016 Danish outbreaks and those
226 from Poland and Holland. Therefore, the dataset was split at the root thus forming two
227 separate alignments: A) containing the main Saeby cluster (A, fig. 2), and BC)
228 containing the remaining samples (B and C, fig. 2). Subsequent analyses using
229 TempEst yielded small residuals for the A-cluster (fig. 1), indicating a strict
230 molecular clock was appropriate to describe the evolution of this set of sequences,
231 while the residuals in the BC-cluster (fig. 1) were larger and non-linearly related,
232 suggesting either that the rates had changed through time, the use of a relaxed clock,
233 or that data was unsuitable for clock models (Rambaut et al. 2016). Sequences with
234 large y-axis deviation from the regression line, indicating problems with the sequence
235 (e.g. low base quality, bad assembly), were removed from the alignment.

3.2. Phylogenetic results

3.2.1. Estimating the molecular clock rate

The phylogenetic relationships were inferred in the Bayesian framework implemented in Beast2 v.2.4.4 (Bouckaert et al. 2014) applying an HKY-model with an L-shaped Gamma-rate distribution accounting for a number of invariable sites, a strict molecular clock, and a coalescent exponential tree prior.

The root and the branches dividing Saeby-cluster from the sequences isolated from the 2015/2016 Danish outbreaks and those from Poland and Holland was very deep (fig. 2), and since this might influence the ability to accurately infer the clock-rate, the dataset was split at the root into two new alignments, as described above. The phylogenetic analysis was repeated for each dataset (fig. 3). Cluster A (Saeby) exhibited a declining exponential growth, while the BC-dataset was best described with a constant population parameter (table 1). The mean clock-rates for the models best fitting cluster A and BC were in the range of 3×10^{-4}-3.4×10^{-4} and 1.5×10^{-4} - 8.4×10^{-3} substitutions per site per year, respectively. These rates are similar to those reported for other parvoviruses, such as Canine parovirus-2 (Shackelton et al. 2005) and Porcine parvovirus (Streck et al. 2015). An overview of clock-rate estimates, growth rate parameters, and corresponding 95% HPD's are provided in table 1. The more diverse BC-cluster also had a clock-rate, which could be a reflection of a more aggressive viral strain, a more susceptible animal population, a more adapted viral strain, or a perceived higher clock-rate due to sampling over a short time-span.

3.2.2. Viral population diversity and evolution through time

Bayesian skyline plots were generated for each cluster in order to investigate their epidemiological histories and evolutionary dynamics over time. The effective population size of the Danish Saeby strain had remained constant until the beginning of the 2000's, where it sharply decreased (fig. 4), likely the result of that the Danish AMDV control programme became regulated by law in 1999 (Anon 2009). The flat skyline prior to 1999, suggested that the viral population had been stable through years, and that the rate of removal of infected animals was somewhat equal to the number of new cases. However, this could also be an artefact, since the coalescent

skylines are known to struggle with estimating population sizes further backwards in time when there are fewer samples (Ho & Shapiro 2011). The mean effective population size was larger within the BC cluster compared to in the A cluster, but with a wider 95% HPD's (fig. 4), possibly reflecting the uncertainties related to these data being sampled from a shorter time-span. The overall appearances of the skyline graphs for each cluster were similar regardless if the entire datasets or one sample per farm was analysed (data not shown).

3.3. Sliding window phylogenetic analysis

Shorter DNA fragments are easier to PCR-amplify and are better suited for routine diagnostic purposes than whole genome sequences. Therefore, a sliding-window analysis was performed to investigate if there was a genomic sub sequence that could provide similar phylogenetic resolution as the full genome. The relative phylogenetic resolutions for each of the 400-bp partitions compared to the full genome were in the range of 18% (nt. 3751) to 43% (nt. 1751 and 1776) (fig. 5). The MCC-trees for the entire genome and the partitions starting at nt 1751, 3751, and 501 (closest to the partial NS1 gene position) revealed inconsistent topologies (fig. 6), suggesting that for AMDV, there is no single genomic sub region providing similar phylogenetic resolution as the entire genome.

3.4. Selection pressure

The regulatory protein NS1 has an important role in parvovirus replication during infection (Christensen et al. 1995; Gottschalck et al. 1994), while the viral capsid protein VP2 has been linked to determination of host range and pathogenicity (Bloom et al. 2001; Bloom et al. 1998). For the NS1 gene we found *overall* dN/dS ratios of 0.64 and 0.59 for cluster A and BC, respectively, and 0.23 (A) and 0.25 (BC) for the VP2 gene, perhaps suggesting the evolution of these genes mainly was driven by negative selection or indicating the challenge to detect selection pressure when there is high sequence homology (Posada et al. 2002).

Looking at the selection at specific sites in the NS1 gene, dominant negative selection was evident on four (cluster A) and 25 (cluster BC) codons, while positive selection was acting on three (A) and two (BC) codons (fig. 7). For the VP2 gene, 15 (A)

302 respectively 62 (BC) codons were negatively selected, and positive selection was
303 acting on a single codon for cluster A and on five in cluster BC (fig. 7). To investigate
304 if the differences between the clusters could be assigned to the different sampling
305 time-spans (i.e. cluster BC was sampled during two seasons, while cluster A was
306 spread over 10 years), the dN/dS analyses were repeated for cluster A reduced to
307 contain only sequences from 2014 and 2015, and since it remained in a similar range
308 for both genes (data not shown), this indicates that the evolution of the Saeby strain
309 was driven mainly by negative selection.
310

311 ## 3.5. Amino acid changes

312 A selection of previously described sites were examined (table 2) to explore if the
313 suggested *in vivo* differences in pathogenicity between strains in the two clusters
314 could be linked to individual genomic regions and/or amino acid (aa) changes.
315

316 ### 3.5.1. Replication factors

317 The GKRN-region between NS1 aa 421-492 (Gottschalck et al. 1994), suggested to
318 be important for the ATP-ase function and for promoting conformational changes in
319 mRNA affecting translation, and the promoters, P3 (nt 151-160) assumed to initiate
320 transcription of all mRNA (Bloom et al. 1988), and P36 (around nt 1744) (Qiu et al.
321 2006; Bloom et al. 1988), were both conserved between the clusters (fig. 8 and 10).
322 The PKR-region between NS1 aa 623-625, previously reported to be involved in
323 nuclear localisation (Gottschalck et al. 1994) was also conserved, but subject to
324 negative selection in cluster BC (table 2). This high level of conservation likely
325 indicates the importance of these regulatory factors for AMDV replication.
326

327 ### 3.5.2. Pathogenicity and host-selection markers

328 The VP2 N-terminus have previously been suggested to influence host-range and
329 AMDV-G's ability to grow in Crandell feline kidney (CRFK) cells (Bloom et al.
330 1998). In the present study, several sites in this region were under negative selection
331 (fig. 7) and differed on aa level between the clusters, also between the less aggressive
332 strains (Saeby, AMDV-G) and the presumed more pathogenic AMDV-Utah and the
333 outbreak related BC-strains (e.g. VP2 aa 115, table 2).
334

Comparisons of other VP2 gene sites with suggested influence on pathogenicity, such as the hypervariable region VP2 aa 231-242 (Oie et al. 1996), VP2 aa 395 and 534 (Bloom et al. 1998), were mainly subjective to negative selection and exhibited little variation both within and between the clusters (fig. 9). Amino acid 420 has been proposed to increase viral fitness by prevention of caspase cleavage (Cheng et al. 2010), and in agreement with previous studies (Sang et al. 2012; Oie et al. 1996; Bloom et al. 1988; Hagberg et al. 2016) this site was conserved between the strains isolated here. Several amino acids between VP2 428-448, suggested to define AMDV host-range (McKenna et al. 1999; Wang et al. 2014), were under negative selection (fig. 7) and highly conserved at aa level (fig. 8 and 9) in the present study, possibly confirming their importance for pathogenesis.

4. Discussion

Sequence analyses in a phylogenetic framework is a useful tool both for confirming known patterns of spread and revealing potential new links suggested by epidemiological data (Valdazo-González et al. 2012; Metzker et al. 2002; Hagberg et al. 2017). By comparison of multiple shorter partitions across the AMDV genome, we showed that none of these segments provided nearly as good phylogenetic resolution as the entire genome. However, despite this fact, even the genome sequence data needs to be supplemented with meta-data such as sampling-dates or farm prevalence in order to more accurately date events, especially when the sequence diversity is low (Leekitcharoenphon et al. 2014; Hagberg et al. 2017).

Using whole genome sequences isolated from AMDV field strains we reconstructed the phylogenetic relationships and described the dynamics of the virus in Denmark during the past ten years. We showed that the strains clustered in two distinct groups separated by a deep split, suggesting there was a large number of unsampled hosts and that the populations originated from two unrelated strains that have evolved independently from each other. In order to investigate the dynamics *within* each cluster, we divided the dataset according to this split: cluster BC mainly originating from a recent outbreak, and cluster A representing years of sampling from a persistent population. Furthermore, this is the first study to provide robust estimate of AMDVs evolutionary rates based on the entire viral genome, and to show these rates are in line

with those reported for other parvoviruses with veterinary importance, such as Canine parovirus-2 (Shackelton et al. 2005) and Porcine parvovirus (Streck et al. 2015). The genetic data presented here supports the previous finding (Christensen et al. 2011) that despite this elevated evolutionary rate, the effective AMDV population size was reduced *after* the legal regulation of the Danish control programme in 1999, supporting its efficiency in reducing the viral load through consistent removal of infected animals and thereby viral strains from the population.

The difference between the numbers of sites under selection between strains in the two clusters was striking: in cluster BC a much larger proportion of sites were subject especially to negative selection, perhaps reflecting that some of these strains were sampled in an early phase epidemic (the Danish sequences) and in Poland and Holland, countries with little or no control measures towards AMDV, and thus were better adapted to their hosts and environments. The Danish Saeby strain (cluster A) on the other hand, represents a population subject to years of thorough control measures, and one would expect that the continuous removal of infected animals from the farms would contribute to positive selection for viral strains with e.g. more efficient replication or better ability to resist environmental factors. The regulatory regions were highly conserved both within and between the clusters, and in the BC cluster strains there was a negative pressure on the nuclear localisation function, perhaps suggesting the importance of preserving these factors for the viability of the virus. The overall replication pressure was negative in both clusters, suggesting the non-structural gene evolved under a similar type of pressure as the viral capsid. This could indicate that humoral immune selection was not the primary driver of the high evolutionary rates in AMDV, at least not for the VP genes, a finding which is in concordance with studies in e.g. the related Human parvovirus B19 (Shackelton & Holmes 2006). Instead, factors such as efficient replication or an increased capsid stability, and thus also increased ability to survive in the environment, might be more important drivers for AMDV evolution. Having this knowledge of the degree of conservation of individual sites and potential pathogenic markers between the strains could be useful e.g. for designing sensitive and discriminatory diagnostics tests or potentially to adjust outbreak-related control measures according to strain characteristics.

Another important evolutionary driver is recombination, which will be revealed in a phylogenetic study by different parts of a sequence having different closest ancestors. Previous studies of AMDV (Canuti et al. 2016; Hagberg et al. 2017) and of other parvoviruses (Shackelton et al. 2007; Streck et al. 2011) have shown incongruent phylogenies depending on which region of the genome were used for the inferences. E.g. a set of full-length strains from Newfoundland clustered differently when the phylogeny was built using NS1, VP2, and even on different parts of the genes (Canuti et al. 2016), and the authors suggested that recombination could have influenced this inconsistency. We were not able to detect recombination in the data presented here, but it should be kept in mind recombination detection has low power when there is little sequence divergence (Posada et al. 2002). Prior to this study, there were only a few publically available AMDV whole genome sequences, and thus the entire genomic diversity on a global scale is fairly unknown. An interesting thought and question for future studies to address, is whether denser sampling could have revealed any links between the two clusters.

In summary, this is the first study to investigate the evolution and selection pressure in a comprehensive set of full-length AMDV field strains. Using a Bayesian skyline analysis we show that the genetic data supported the efficiency of the Danish control programme and we showed that a routinely applied stamping-out strategy could reduce the genetic diversity in an AMDV population. We furthermore present robust evolutionary rate-estimates for AMDV, which we suggest should be used to more accurately determine the timing and the origin in future outbreaks.

Acknowledgements

Co-workers from the diagnostic and farmers relations departments in Kopenhagen Fur are gratefully thanked for collecting samples and for providing valuable background information about the farms history. Haiko Koenen (DVM, DAC Zuidoost) is gratefully thanked for collecting samples in The Netherlands. The laboratory technician Sari Mia Dose at the Danish National Veterinary Institute is gratefully thanked for preparing more than half of the samples for sequencing. Anne-Sofie Hammer, Copenhagen University, is acknowledged for providing tissue material from

two wild mink originating from Bornholm.

This work was supported by the Research Foundation of the Danish Fur breeder's Association" and the "Innovation Fund Denmark" (award number: 16479). None of the sponsors had influence on study design, interpretation of results, or the decision to publish.

All sample materials were obtained from freezer archives or from animals euthanized for other purposes than this particular study.

References

Anon, 2009. *Danish Executive Order 1447 of 15/12/2009*, Available at: https://www.retsinformation.dk/Forms/R0710.aspx?id=129366 [Accessed May 21, 2015].

Bloom, M.E. et al., 1998. Construction of pathogenic molecular clones of Aleutian mink disease parvovirus that replicate both in vivo and in vitro. *Virology*, 251(2), pp.288–96. Available at: http://www.sciencedirect.com/science/article/pii/S0042682298994260 [Accessed May 13, 2015].

Bloom, M.E. et al., 2001. Identification of aleutian mink disease parvovirus capsid sequences mediating antibody-dependent enhancement of infection, virus neutralization, and immune complex formation. *Journal of virology*, 75(22), pp.11116–27. Available at: http://jvi.asm.org/content/75/22/11116.full [Accessed January 8, 2015].

Bloom, M.E. et al., 1988. Nucleotide sequence and genomic organization of Aleutian mink disease parvovirus (ADV): sequence comparisons between a nonpathogenic and a pathogenic strain of ADV. *Journal of virology*, 62(8), pp.2903–15. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=253728&tool=pmcentrez&rendertype=abstract.

Bloom, M.E., Race, R.E. & Wolfinbarger, J.B., 1980. Characterization of Aleutian disease virus as a parvovirus. *Journal of virology*, 35(3), pp.836–43. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=288877&tool=pmcentrez&rendertype=abstract [Accessed March 20, 2015].

Bouckaert, R. et al., 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS computational biology*, 10(4), p.e1003537. Available at: http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003537 [Accessed July 11, 2014].

Bouckaert, R., 2015. bModelTest: Bayesian site model selection for nucleotide data. *bioRxiv*.

Canuti, M. et al., 2016. Driving forces behind the evolution of the Aleutian mink

478 disease parvovirus in the context of intensive farming. *Virus Evolution*, 2(1),
479 p.vew004. Available at:
480 http://ve.oxfordjournals.org/lookup/doi/10.1093/ve/vew004.
481 Cheng, F. et al., 2010. The capsid proteins of Aleutian mink disease virus activate
482 caspases and are specifically cleaved during infection. *Journal of virology*, 84(6),
483 pp.2687–96. Available at:
484 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2826067&tool=pmce
485 ntrez&rendertype=abstract [Accessed March 20, 2015].
486 Christensen, J. et al., 1995. Purification and characterization of the major
487 nonstructural protein (NS-1) of Aleutian mink disease parvovirus. *Journal of*
488 *virology*, 69(3), pp.1802–9. Available at:
489 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=188788&tool=pmcen
490 trez&rendertype=abstract.
491 Christensen, L.S. et al., 2011. Diversity and stability of Aleutian mink disease virus
492 during bottleneck transitions resulting from eradication in domestic mink in
493 Denmark. *Veterinary microbiology*, 149(1–2), pp.64–71. Available at:
494 http://www.ncbi.nlm.nih.gov/pubmed/21112164 [Accessed May 21, 2014].
495 Decaro, Nicola et al., 2012. 12. Parvovirus infections. In D. Gavier-Widén, J. P. Duff,
496 & A. Meredith, eds. *Infectious Diseases of Wild Mammals and Birds in Europe*.
497 Oxford, UK: Wiley-Blackwell, pp. 181–285.
498 Gottschalck, E. et al., 1994. Sequence comparison of the non-structural genes of four
499 different types of Aleutian mink disease parvovirus indicates an unusual degree
500 of variability. *Archives of Virology*, 138(3–4), pp.213–31. Available at:
501 http://www.ncbi.nlm.nih.gov/pubmed/7998830 [Accessed May 27, 2014].
502 Hagberg, E.E. et al., 2016. A fast and robust method for whole genome sequencing of
503 the Aleutian Mink Disease Virus (AMDV) genome. *Journal of Virological*
504 *Methods*. Available at:
505 http://www.sciencedirect.com/science/article/pii/S0166093415300343.
506 Hagberg, E.E. et al., 2017. Evolutionary analysis of whole genome sequences from
507 Aleutian Mink Disease Viruses confirms inter-farm transmission. *Journal of*
508 *General Virology*.
509 Ho, S.Y.W. & Shapiro, B., 2011. Skyline-plot methods for estimating demographic
510 history from nucleotide sequences. *Molecular Ecology Resources*, 11(3),
511 pp.423–434.
512 Jensen, T.H. et al., 2011. Implementation and validation of a sensitive PCR detection
513 method in the eradication campaign against Aleutian mink disease virus. *Journal*
514 *of virological methods*, 171(1), pp.81–5. Available at:
515 http://www.ncbi.nlm.nih.gov/pubmed/20951744 [Accessed May 21, 2014].
516 Katoh, K. & Standley, D.M., 2013. MAFFT multiple sequence alignment software
517 version 7: improvements in performance and usability. *Molecular biology and*
518 *evolution*, 30(4), pp.772–80. Available at:
519 http://mbe.oxfordjournals.org/content/30/4/772 [Accessed July 13, 2014].
520 Kearse, M. et al., 2012. Geneious Basic: an integrated and extendable desktop
521 software platform for the organization and analysis of sequence data.
522 *Bioinformatics (Oxford, England)*, 28(12), pp.1647–9. Available at:
523 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3371832&tool=pmce
524 ntrez&rendertype=abstract [Accessed July 10, 2014].
525 Kühnert, D., Wu, C.-H. & Drummond, A.J., 2011. Phylogenetic and epidemic
526 modeling of rapidly evolving infectious diseases. *Infection, genetics and*
527 *evolution : journal of molecular epidemiology and evolutionary genetics in*

infectious diseases, 11(8), pp.1825–41. Available at: http://www.sciencedirect.com/science/article/pii/S156713481100284X [Accessed December 17, 2015].

Leekitcharoenphon, P. et al., 2014. Evaluation of Whole Genome Sequencing for Outbreak Detection of Salmonella enterica J. A. Chabalgoity, ed. *PLoS ONE*, 9(2), p.e87991. Available at: http://dx.plos.org/10.1371/journal.pone.0087991 [Accessed August 15, 2016].

Leimann, A. et al., 2015. Molecular epidemiology of Aleutian mink disease virus (AMDV) in Estonia, and a global phylogeny of AMDV. *Virus research*, 199(2), pp.55–61. Available at: http://www.sciencedirect.com/science/article/pii/S0168170215000179 [Accessed January 26, 2015].

Li, Y. et al., 2012. Genetic characterization of Aleutian mink disease viruses isolated in China. *Virus genes*, 45(1), pp.24–30. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22415541 [Accessed May 21, 2014].

Martin, D.P. et al., 2015. RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evolution*, 1(1), pp.1–5. Available at: http://ve.oxfordjournals.org/cgi/doi/10.1093/ve/vev003.

McKenna, R. et al., 1999. Three-Dimensional Structure of Aleutian Mink Disease Parvovirus: Implications for Disease Pathogenicity. *J. Virol.*, 73(8), pp.6882–6891. Available at: http://jvi.asm.org/content/73/8/6882.abstract [Accessed March 19, 2015].

Metzker, M.L. et al., 2002. Molecular evidence of HIV-1 transmission in a criminal case. *Proceedings of the National Academy of Sciences of the United States of America*, 99(22), pp.14292–7. Available at: http://www.pnas.org/content/99/22/14292.full [Accessed August 3, 2015].

Nituch, L.A. et al., 2012. Molecular epidemiology of Aleutian disease virus in free-ranging domestic, hybrid, and wild mink. *Evolutionary applications*, 5(4), pp.330–40. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3353359&tool=pmcentrez&rendertype=abstract [Accessed March 20, 2015].

Oie, K.L. et al., 1996. The relationship between capsid protein (VP2) sequence and pathogenicity of Aleutian mink disease parvovirus (ADV): a possible role for raccoons in the transmission of ADV infections. *Journal of virology*, 70(2), pp.852–61. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=189888&tool=pmcentrez&rendertype=abstract.

Pond, S. L., Frost, S.D.W., 2005. Datamonkey: Rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics*, 21(10), pp.2531–2533.

Posada, D., Crandall, K.A. & Holmes, E.C., 2002. Recombination in Evolutionary Genomics. *Annual Review of Genetics*, 36(1), pp.75–97. Available at: http://www.annualreviews.org/doi/10.1146/annurev.genet.36.040202.111115 [Accessed September 29, 2016].

Qiu, J. et al., 2006. The transcription profile of Aleutian mink disease virus in CRFK cells is generated by alternative processing of pre-mRNAs produced from a single promoter. *Journal of virology*, 80(2), pp.654–62. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1346859&tool=pmcentrez&rendertype=abstract [Accessed November 4, 2014].

Rambaut, A. et al., 2016. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evolution*, 2(1), pp.1–7.

Rambaut, A. et al., 2014. Tracer v1.6. Available at: http://beast.bio.ed.ac.uk/Tracer.
Ryt-Hansen, P. et al., 2017. *Outbreak investigation of Aleutian Mink Disease Virus (AMDV) using partial NS1 gene sequencing*,
Sang, Y. et al., 2012. Phylogenetic analysis of the VP2 gene of Aleutian mink disease parvoviruses isolated from 2009 to 2011 in China. *Virus genes*, 45(1), pp.31–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22415542 [Accessed May 21, 2014].
Shackelton, L.A. et al., 2007. Comparative analysis reveals frequent recombination in the parvoviruses. *The Journal of general virology*, 88(Pt 12), pp.3294–301. Available at: http://jgv.microbiologyresearch.org/content/journal/jgv/10.1099/vir.0.83255-0#tab2 [Accessed February 1, 2016].
Shackelton, L.A. et al., 2005. High rate of viral evolution associated with the emergence of carnivore parvovirus. *Proceedings of the National Academy of Sciences of the United States of America*, 102(2), pp.379–84. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15626758 [Accessed August 17, 2016].
Shackelton, L.A. & Holmes, E.C., 2006. Phylogenetic Evidence for the Rapid Evolution of Human B19 Erythrovirus. *Journal of Virology*, 80(7), pp.3666–3669. Available at: http://jvi.asm.org/cgi/doi/10.1128/JVI.80.7.3666-3669.2006 [Accessed August 22, 2016].
Stadler, T. et al., 2014. Insights into the Early Epidemic Spread of Ebola in Sierra Leone Provided by Viral Sequence Data. *PLoS Currents*, 6. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4205153&tool=pmcentrez&rendertype=abstract [Accessed January 13, 2015].
Streck, A.F. et al., 2011. High rate of viral evolution in the capsid protein of porcine parvovirus. *Journal of General Virology*, 92(11), pp.2628–2636.
Streck, A.F., Canal, C.W. & Truyen, U., 2015. Molecular epidemiology and evolution of porcine parvoviruses. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases*, 36, pp.300–6. Available at: http://www.sciencedirect.com/science/article/pii/S1567134815300071 [Accessed February 22, 2016].
Su, Y.C.F. et al., 2015. Phylodynamics of H1N1/2009 influenza reveals the transition from host adaptation to immune-driven selection. *Nature Communications*, 6, p.7952. Available at: http://www.nature.com/doifinder/10.1038/ncomms8952.
Valdazo-González, B. et al., 2012. Reconstruction of the Transmission History of RNA Virus Outbreaks Using Full Genome Sequences: Foot-and-Mouth Disease Virus in Bulgaria in 2011 Y. E. Khudyakov, ed. *PLoS ONE*, 7(11), p.e49650. Available at: http://dx.plos.org/10.1371/journal.pone.0049650 [Accessed August 31, 2016].
Wang, Z. et al., 2014. Molecular epidemiology of Aleutian mink disease virus in China. *Virus research*, 184, pp.14–9. Available at: http://www.sciencedirect.com/science/article/pii/S0168170214000604 [Accessed May 30, 2016].

Figures

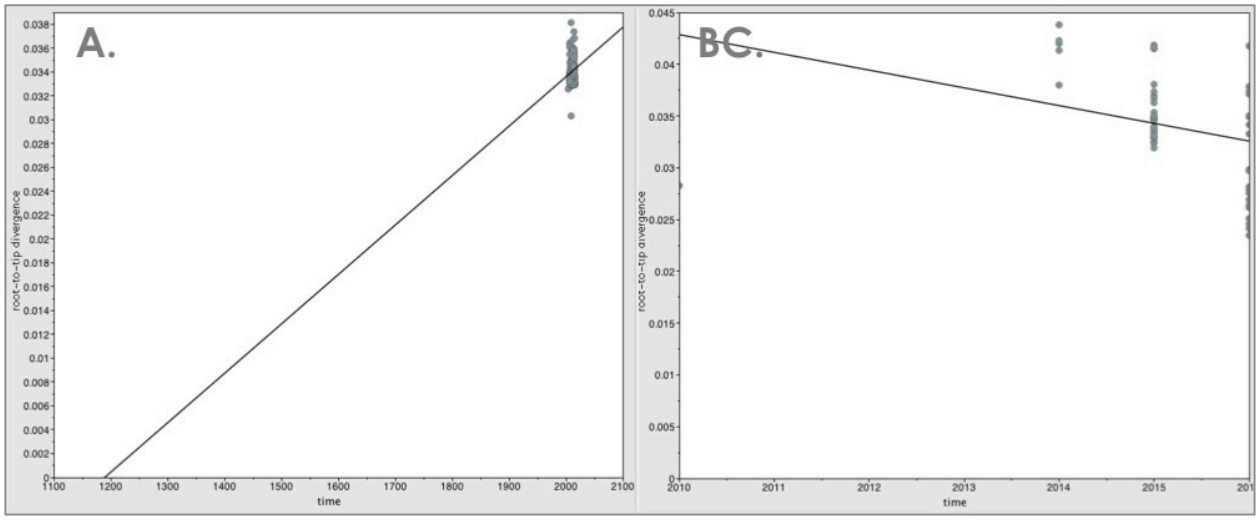


Figure 1. **Investigation of clocklikeness.** Linear regression showing for each sequence the relationship between sampling year (time, x-axis) and its divergence from the root (y-axis), based on ML-phylogenies for each dataset. Note that for the BC-dataset TempEst cannot be used to estimate rates (BC.).

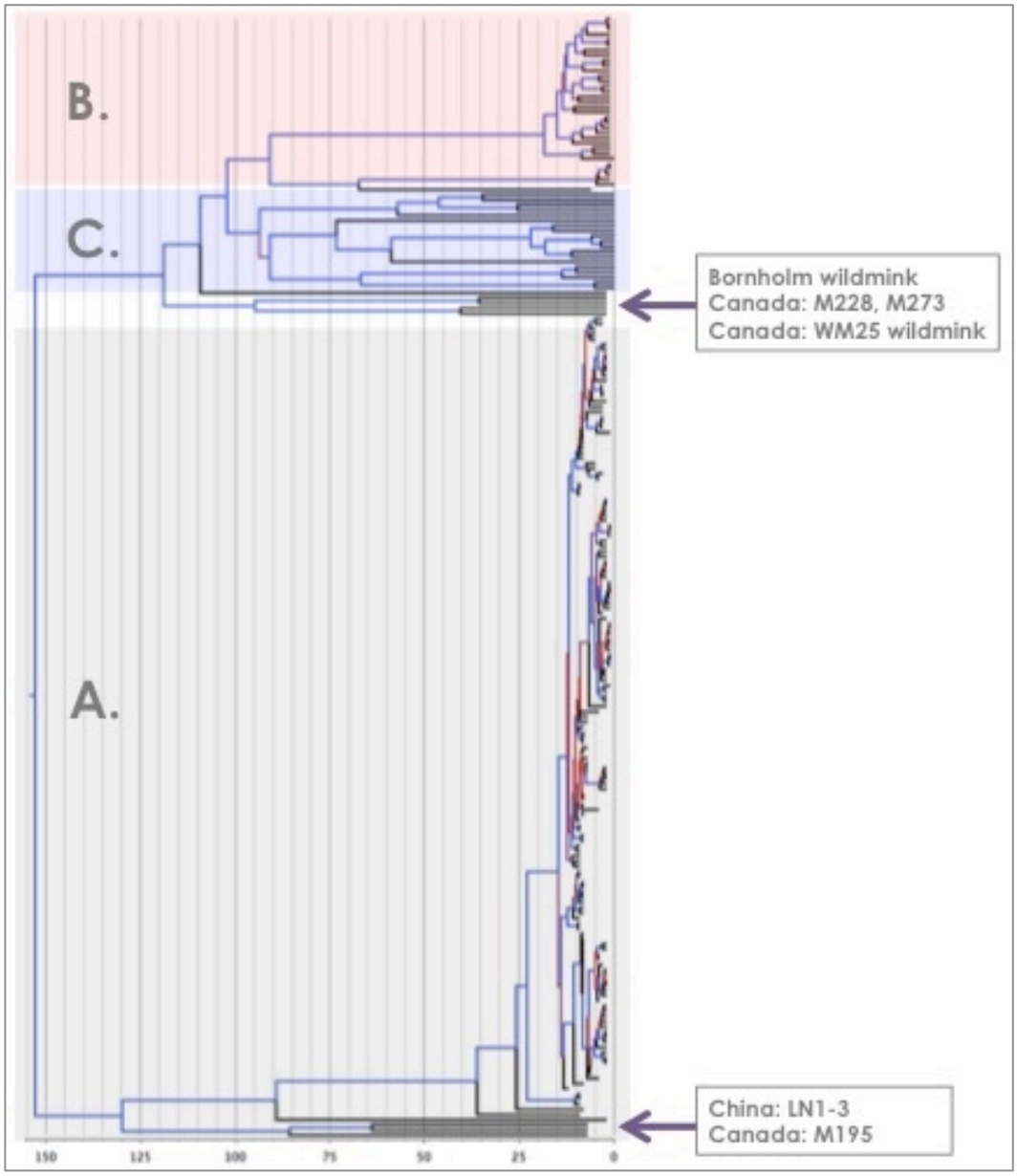


Figure 2. Notice the deep root separating the main Danish cluster, i.e. Saeby, (A) from the Holstebro sequences (B) and those isolated in Holland and Poland 2016 (C). Furthermore it can be seen that the wildmink sequences from Bornholm, the wildmink from Canada 2014, and two of the Canadian sequences cluster together with the BC group, while the Chinese sequences and the remaining Canadian sequence, are closer related to the Saeby cluster (A). The MCC tree was constructed in BEAST2 with an HKY-model, estimating the number of invariable sites and gamma-rate distribution from the data, calibration on sampling years, and applying a strict molecular clock and an exponential coalescent population model. The MCMC chain was run for 50M iterations. Branches are coloured according to their posterior probabilities (Bayesian support values), where red is low and blue is high (range 0.5-1). The X-axis represents time in years.

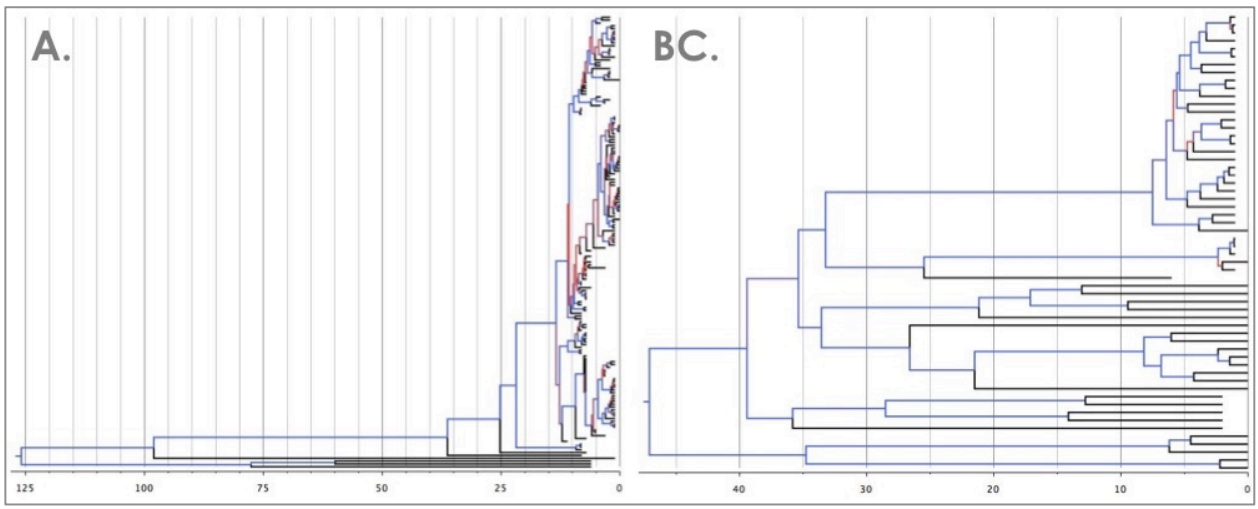


Figure 3. MCC tree for each cluster, branches are coloured according to their posterior probabilities (Bayesian support values), where red is low and blue is high (range 0.5-1). The X-axis represents time in years. The trees were constructed in BEAST2 with an HKY-model, estimating the number of invariable sites and gamma-rate distribution from the data, calibration on sampling years, applying a strict molecular clock and an exponential (A) or constant (B) coalescent population model. The MCMC chains were run for 50M iterations.

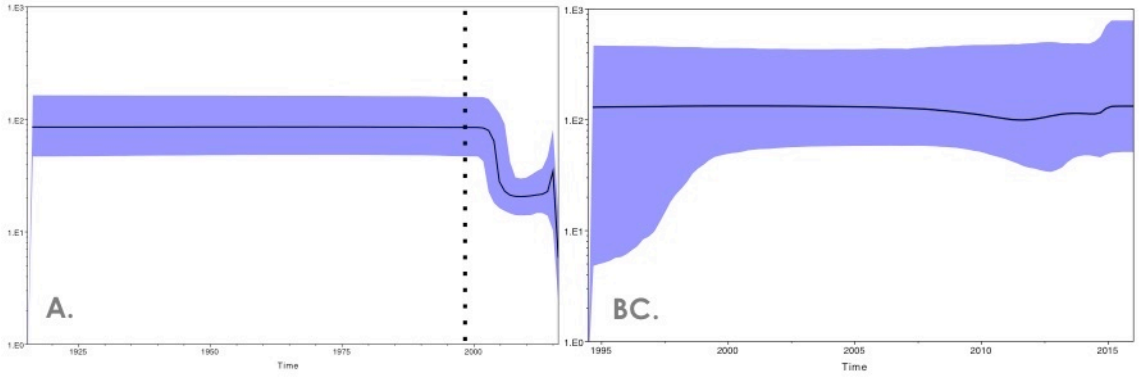


Figure 4. Bayesian Skyline plots for cluster A and cluster BC. The phylogenies were constructed in BEAST2 with an HKY-model, estimating the number of invariable sites and gamma-rate distribution from the data, calibration were on sampling years, and applying a strict molecular clock and a coalescent Bayesian skyline population model. The MCMC chain was run for 50M iterations. The x-axes represent time in years, the y-axes represent the N_{et} (the product of the effective population size and the generation time in years), the thick black line is the median N_{et}, the solid purple field its 95% HPD, and the stippled black line indicate implementation of the Danish control programme in 1999.

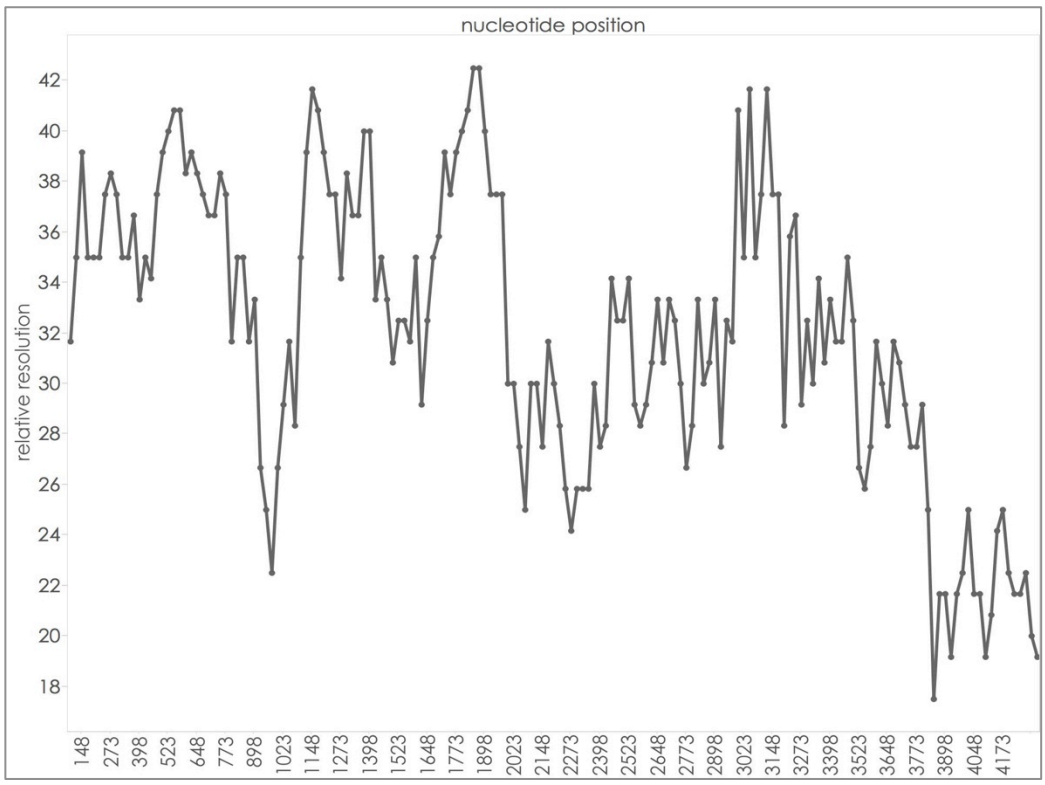


Figure 5. Sliding-window analysis. Bayesian phylogenetic trees were created using 400 bp windows spaced at 25 bp intervals across the whole-genome alignment. The relative resolution was measured using the number of internal nodes (y-axis) as a function of the start-position of the 400 bp window (x-axis) relative to the AMDV-G genome.

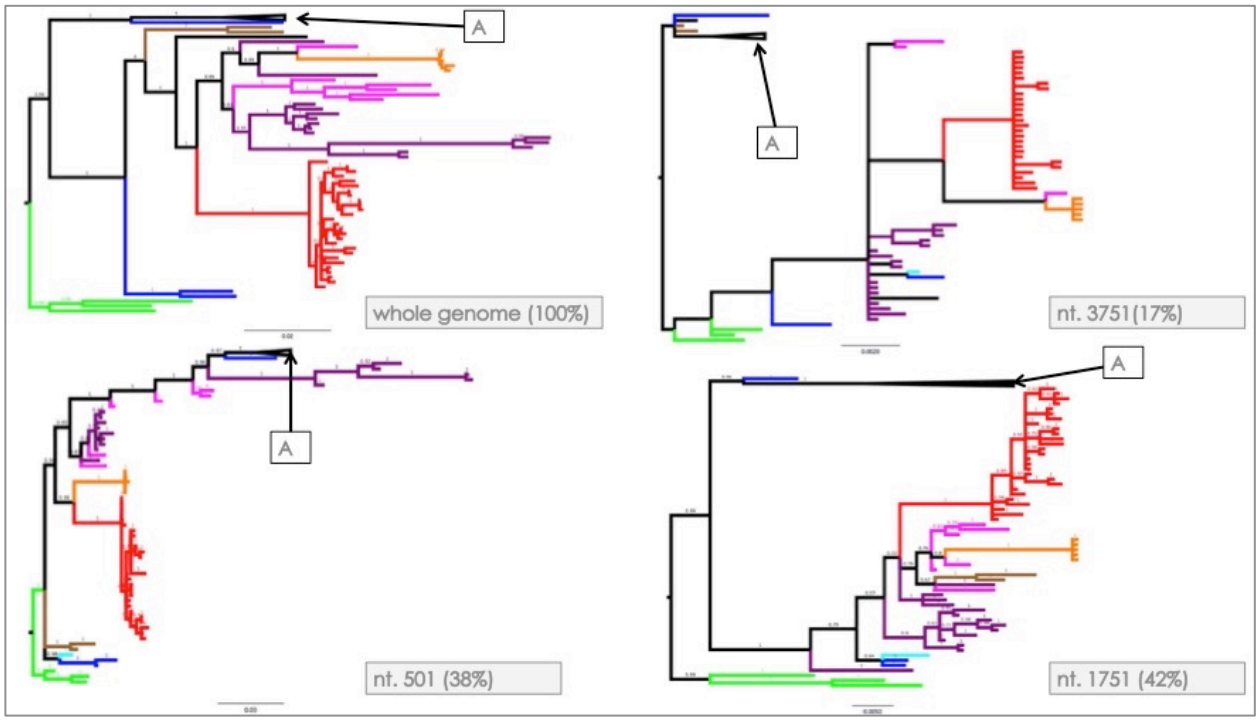


Figure 6. Comparison of different partitions. From sliding-window analysis, whole geome, and 400-bp alignments providing the highest (42%) and lowest (17%) relative

resolution, and the partition correspinding approximately to the partial NS1 gene (38%). To improve visualisation the Sæby cluster (black A) was collapsed, Holstebro sequences (red), Zealand (orange), Holland (pink), Poland (purple), Canada (blue), wildmink Canada (turquise), wildmink Bornholm (brown). Notice the red (Holstebro) sequence, which clusters separately based on the NS1 gene, while together with the other Holstebro sequences based on VP2. MCC trees from phylogenies constructed in MrBayes v.3.2, applying an HKY-model and estimating the number of invariable sites and gamma-rate distributions from the data. The MCMC's were run for 5M iterations. Branch labels represent posterior probabilities for each clade (Bayesian support values), and branch-lengths represent substitutions per site as indicated by the scale bars.

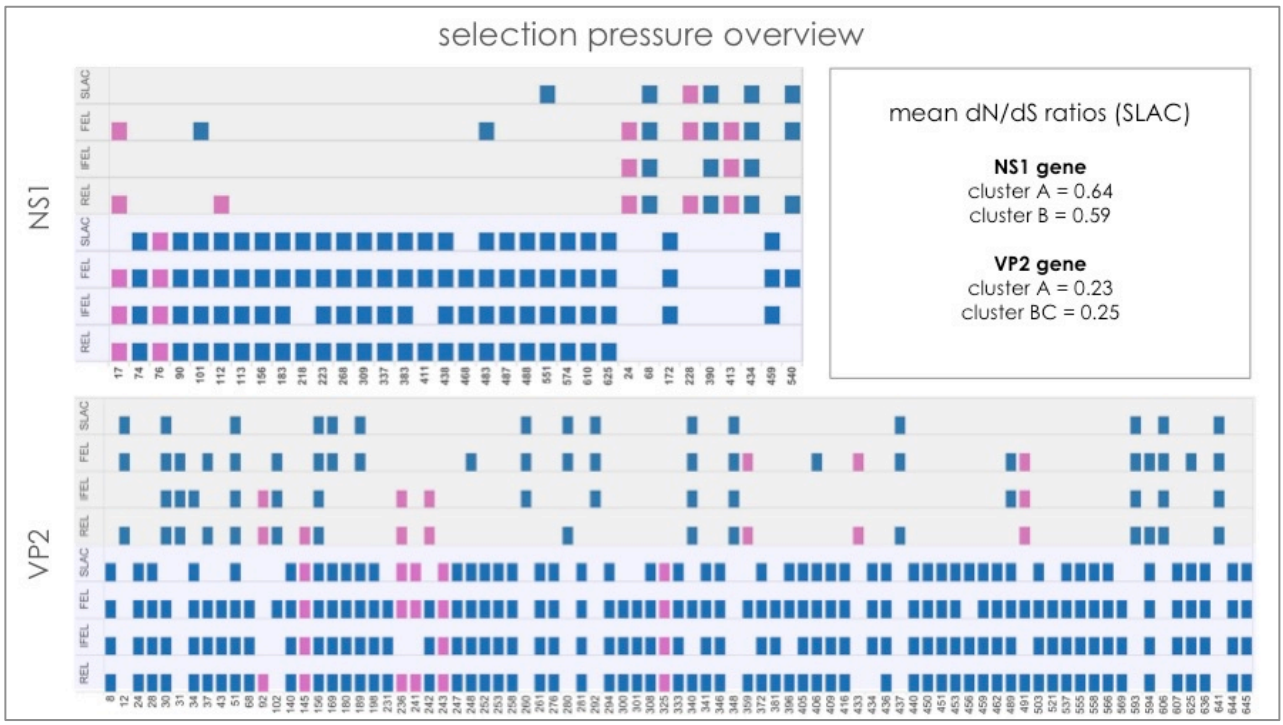


Figure 7. Selection pressure overview. Cluster A (grey shaded fields) and cluster BC (blue shaded fields) for each of the two major genes: NS1 and VP2. The X-axes represent codons: positively selected sites are indicated with pink boxes, and negatively selected sites with blue. The sites were identified using SLAC, FEL, IFEL, and REL at Datamonkey.org, and the mean dN/dS ratios were calculated using SLAC.

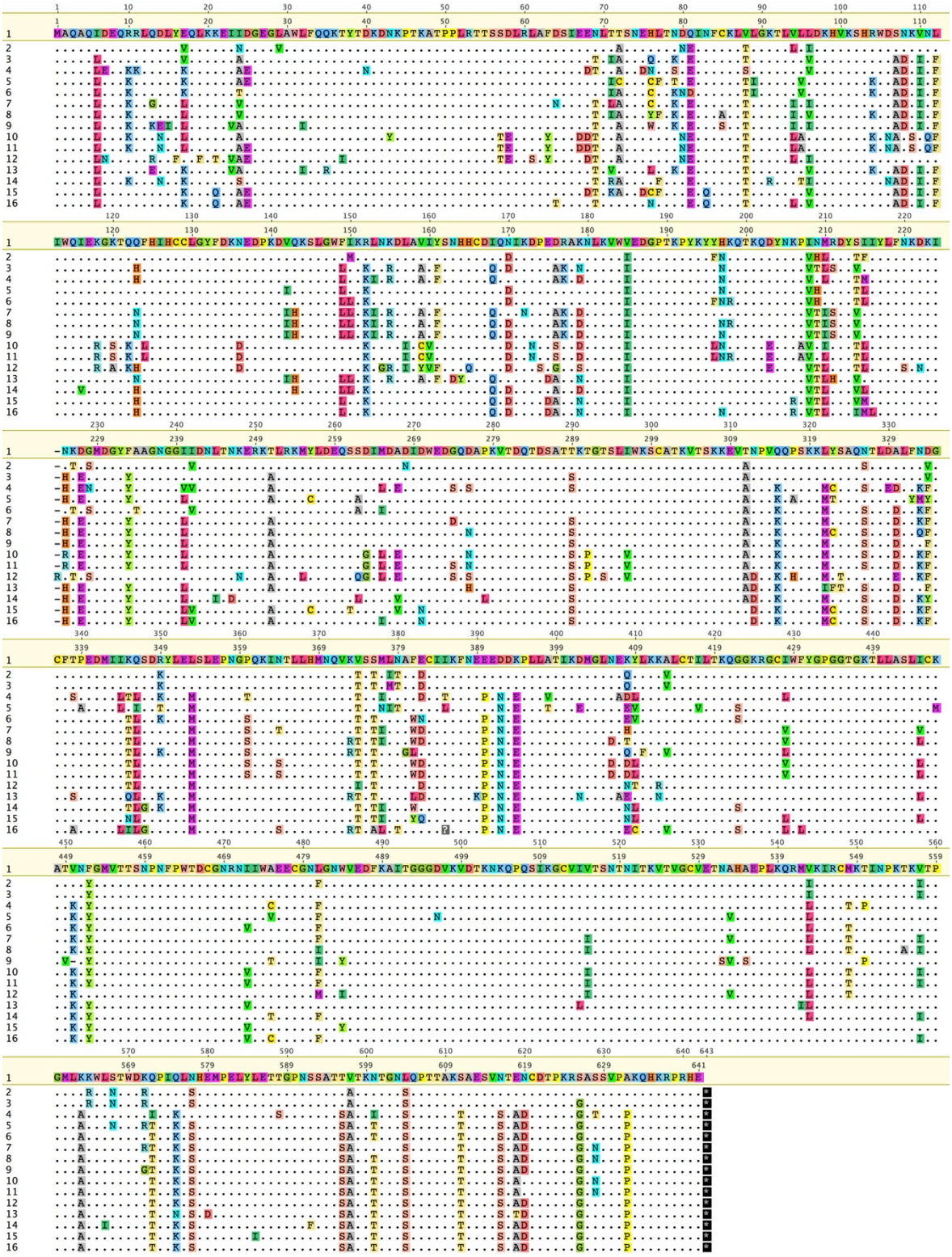


Figure 8. NS1 gene. Amino acid alignment of the NS1 gene for a consensus of cluster A (1), AMDV-G (2), AMDV-Utah (3), and a selection of sequences from cluster BC (4-16). Figures created in Geneious v.7.1.5.

Figure 9. VP2 gene. Amino acid alignment of the VP2 gene for a consensus of cluster A (1), AMDV-G (2), AMDV-Utah (3), and a selection of sequences from cluster BC (4-16). Figures created in Geneious v.7.1.5.

Tables

Table 1. Clock-rate estimates. Overview of the clock rate-estimates for datasets investigated in this study. Entire dataset n=206 (ABC), cluster A, cluster B, strict-clock (sc), coalescent exponential population prior (cep), coalescent constant population prior (ccp), 95% credibility intervals of the mean values (HPD).

Table 2. Overview of codons. The table provides an overview of selected interesting codons, where there were amino acid (aa) differences between cluster A and BC and/or sites with selection pressure confirmed by other studies. Suggested functions of the sites or regions in question are indicated.

Table 1. Clock-rate estimates

	sc cep				sc ccp		sc cbs	
	growth.rate		clock.rate		clock.rate		clock.rate	
dataset	mean	95% HPD	mean	95% HPD	mean	95% HPD	mean	95% HPD
ABC	-0.0094	[-0.0176, -0.0019]	4,90E-04	[3.8442E-4, 5.9813E-4]	4,96E-04	[3.6936E-4, 5.6885E-4]	-	-
A	-0.025	[-0.0404, -0.0088]	3,63E-04	[2.8843E-4, 4.3727E-4]	na	na	3,65E-04	[2.9423E-4, 4.405E-4]
BC	0.0355	[-0.0003, 0.0738]	1,55E-03	[6.2979E-4, 2.5128E-3]	1,47E-03	[5.6185E-4, 2.4918E-3]	1,36E-03	[3.5279E-4, 2.4093E-3]

Table 2. Overview of codons

NS1	aa change			selection pressure			
aa	cluster A	cluster BC	others'	cluster A	cluster BC	others'	suggested functions
6	I (AMDV-G)	L (Utah)					
76					pos	pos (Canuti et al., 2016)	ATP-binding site
101					neg	neg (Canuti et al., 2016)	
107	S (AMDV-G)	A (Utah)					
108	N (AMDV-G)	D (Utah)					
112					neg	neg (Canuti et al., 2016)	
438					neg	neg (Canuti et al., 2016)	GKRN-region, ATP-bindning pocket (435-440)
459					neg	neg (Canuti et al., 2016)	GKRN-region
483					neg	neg (Canuti et al., 2016)	GKRN-region
488					neg	neg (Canuti et al., 2016)	GKRN-region
625					neg	neg (Canuti et al., 2016)	PKR-region (623-625)
VP2	aa change			selection pressure			
aa	cluster A	cluster BC	others'	cluster A	cluster BC	others'	function
6	A (AMDV-G, Utah)	D					Host range (1-220)? (Bloom et al. 1998)

8					neg	neg (Canuti et al., 2016)	Host range (1-220)? (Bloom et al. 1998)
28					neg	neg (Canuti et al., 2016)	Host range (1-220)? (Bloom et al. 1998)
92	H (AMDV-G, Utah)	A	(Hagberg et al. 2016)				Host range (1-220)? (Bloom et al. 1998)
94	Q (AMDV-G, Utah)	K					Host range (1-220)? (Bloom et al. 1998)
115	Y (AMDV-G)	F (Utah)				neg (Canuti et al., 2016)	Host range (1-220)? (Bloom et al. 1998)
140	D	E (AMDV-G, Utah)			neg	neg (Canuti et al., 2016)	Host range (1-220)? (Bloom et al. 1998)
180					neg	neg (Canuti et al., 2016)	Host range (1-220)? (Bloom et al. 1998)
231					neg	neg (Canuti et al., 2016)	Hypervar 231-242
232	V (AMDV-G)	D (Utah)			neg	neg (Canuti et al., 2016)	Hypervar 231-242
233	A (AMDV-G)	G (Utah)					Hypervar 231-242
242					neg	neg (Canuti et al., 2016)	Hypervar 231-242
282	N (AMDV-G, Utah)	K	(Wang et al., 2014)				
300					neg		CPV host range (Allison et al., 2014)
308					neg	pos (Canuti et al., 2016)	
317	K (AMDV-G, Utah)	L	(Wang et al., 2014)		neg		
323	L (AMDV-G, Utah)	I					CPV host range (Allison et al., 2014)
324	I (AMDV-G, Utah)	T					
327	Q (AMDV-G, Utah)	E (50%)	(Wang et al., 2014)				
359	Y (AMDV-G, Utah)	M					
405					neg	neg (Canuti et al., 2016)	
416					neg	neg (Canuti et al., 2016)	
434	H	H	(Hagberg et al. 2016)		neg		Host range? (McKenna et al., 1999)
437				neg			Caspase cleavage (McKenna et al. 1999)
440					neg		Caspase cleavage (McKenna et al. 1999)
443							
448	T (AMDV-G, Utah)	S/T	(Wang et al., 2014)				Immunepathogenesis?
491	D (Utah)	E	(Hagberg et al. 2016)	pos			Host range? (McKenna et al., 1999)
516	Q (AMDV-G, Utah)	L					
534	D	D					D in pathogeneic strains (Bloom et al. 1998)
544	N (AMDV-G, Utah)	S					

	non-polar uncharged	S-Serine, T-Threonine, N-Aspargine, Q-Glutamine
	polar + (basic)	R-Arginine, H-Histidine, K-Lysine
	polar - (acidic)	D-Aspartic acid, E-Glutamic acid
	hydrophobic	A-Alanine, V-Valine, I-Isoleucine, L-Leucine, M-Methionine, F-Phenylaline, Y-Tyrosine, W-Tryptophan
	special	C-Cysteine, U-Selencysteine, G-Glycine, P-Proline

4.4. MANUSCRIPT 4

Development of a real-time PCR assay for detection of Aleutian Mink Disease Virus.

Status: the project is still in process, and the preliminary results are written in the form of a manuscript draft.

(page numbers are relative to paper)

Development of a real-time PCR assay for detection of Aleutian Mink Disease Virus (AMDV)

Emma E. Hagberg[a,b,*], Anders Krarup[a], Charlotte Hjulsager[c], Carina B. Folsing[a], Anders G. Pedersen[b], Lars E. Larsen[c],

[a]Kopenhagen Diagnostics, Kopenhagen Fur, Glostrup, Denmark
[b]Department of Bioinformatics, Technical University of Denmark, Lyngby, Denmark
[c]National Veterinary Institute, Technical University of Denmark, Frederiksberg, Denmark

*Corresponding author
Emma E. Hagberg

Abstract

This paper describes the status of a hydrolysis probe based real-time PCR assay for detection of AMDV.

Introduction

Aleutian Mink Disease (AMD), sometimes referred to as Plasmacytosis, is worldwide the most important disease in the mink farming industry. The disease affects mink of all ages and is caused by Aleutian Mink Disease Virus (AMDV), a single stranded DNA virus belonging to the family *Parvoviridae* (Bloom et al. 1980) genus *Amdoparvovirus* species *Carnivore amdoparvovirus 1*. Infection results in a harmful activation of the immune system leading to hypergammaglobulinaemia and systemic vascular diseases where infected animals either die due to organ failure, or become persistently infected carriers transmitting the virus within and between herds (Decaro, Nicola et al. 2012). In Denmark AMDV is monitored by a mandatory national control program (Anon 2009), which briefly requires all farms to conduct screening of their animals at regular intervals according to the disease status of the region. This testing

is performed using a fully automated automated ELISA (Dam-Tuxen et al. 2014), which is suitable for high-throughput screening but serology has limitations during early disease stages where the circulating antibody levels either are absent or below the detection limit (Jensen et al. 2015).

Like other parvoviruses AMDV replicates only in dividing cells where it utilizes the host cell's transcription machinery (Fields et al. 2007). AMDV consists of two large open reading frames (ORF's); the left ORF (nucleotide 116-1975) coding for the non-structural (NS) proteins involved in gene regulation and replication, and the right ORF (nucleotide 2241-4346) coding for the viral capsid proteins (VP), and three smaller central ORF's (Alexandersen et al. 1988; Bloom et al. 1988; Hagberg et al. 2016).

Since 2011 the National Veterinary Institute (Frederiksberg, DK) has routinely performed a confirmatory Sanger sequencing of the partial NS1 gene (Jensen et al. 2011) in AMDV positive samples, and the by far most common genotype based on this partial NS1 gene sequencing has been the so-called Saeby strain (Saeby/DEN) (Christensen et al. 2011). However, during the mink farming season of 2015/2016, two new AMDV strains were detected on Danish farms (Ryt-Hansen, Hjulsager, et al. 2017; Ryt-Hansen, Hagberg, et al. 2017). These strains were named according to the regions they first were isolated from, "Holstebro" and "Zealand", and were, based on partial NS1 gene sequencing, deviating from the in Denmark circulating Saeby strain.

The PCR-primers flanking the partial NS1 gene region (Jensen et al. 2011) were designed based on the at that time single available whole genome AMDV isolate: the cell-culture adapted AMDV-G strain (Bloom et al. 1988). Since then, the molecular diagnostic methods have been improved and allow for faster and easily automatable workflows, compared to endpoint PCR, which is labour intensive, relies a subjective gel electrophoresis step for readout, and a time-consuming confirmatory sequencing step for defining the genotype. Thus, an updated tool was needed for easily discriminating between the three main genotypes in order to quickly conclude weather a farm was infected with the "regular AMDV strain" (i.e. Saeby) or a new emerging genotype.

Therefore, in order to ensure sensitive detection of all circulating viral strains and to improve the objectivity and throughput of PCR in the diagnostic pipeline, a new diagnostic tool was desired. And the aim of this study was to develop a fast and specific high-through-put real-time PCR assay to be used as a supplement to the existing serological ELISA testing and ensure detection of AMDV.

Material and methods

Sample material and DNA-extraction

To establish the method AMDV DNA was isolated from the spleens of euthanized mink submitted for diagnostic purposes. The non-virulent strain AMDV-G (cell culture isolate, passage 10) was obtained from The Research Foundation of the Danish Fur Breeders' Association / Antigen Laboratory (Glostrup, DK). Total DNA was extracted using the QIAmp® MinElute Virus Spin Kit (Qiagen, Hilden, D) according to the manufacturer's instructions, with the final DNA elution performed with 50µL low TE-buffer.

Primer and probe design

The PCR primer- and probe sequences listed in table 1 were designed using the Primer3 software (Ye et al. 2012) implemented in Geneious v.7.1.5 (Kearse et al. 2012) and a sequence alignment including whole genomes of AMDV strains of the to date known genotypes (Hagberg 2017; Hagberg, E.E., Larsen, L.E., Krarup, A., Pedersen 2017). It was essential for the assay to equally well detect all known AMD viral strains and therefore four different genomic regions were screened (fig. 1), and for each region three slightly overlapping primer-pairs were designed. In each of these regions a primer-matrix with all possible combinations of forward and reverse primers was ran using SYBR®-green chemistry and the two best performing primer-pairs, assed by gel electrophoresis, were selected for final bench-marking together with their corresponding probes.

SYBR®-green was chosen for the screening due to its lower cost and the additional analytics from running a melting-curve and gel electrophoresis. The melting-curve step was composed of 5 ^{o}C incremental increases in temperature in the range of 65-95

°C. In addition to running the melting curve PCR products were analysed on 0.8% agarose gels stained with ethidiumbromide to confirm correct size and absence of excessive amounts of primer-dimer or unspecific amplification.

The PCR-products were purified according to the manufacturer's instructions using the QIAquick® PCR Purification Kit (Qiagen, Hildren, DE) and submitted for Sanger-sequencing at LGC Genomics (Berlin, DE) to confirm amplification of the correct AMDV strains. The final primer-pair is reported in table 1.

Real-time PCR cycling conditions

Primer-screening PCR reactions were setup as follows: 10 µL Bio-Rad SsoAdvanced™ Universal SYBR® Green Supermix (2X) (cat.no. 1725270, Bio-Rad Laboratories Inc., Hercules, CA), 200 µM primer F and R respectively, 5 µL DNA template, and distilled water up to a total sample volume of 20 µL. The final probe PCR reactions were setup as follows: 10 µL Bio-Rad SsoAdvanced™ Universal Probe Supermix (2X) (cat.no. 1725280, Bio-Rad Laboratories Inc., Hercules, CA), 200 nM primer F and R respectively, 150 nM probe, 5 µL DNA template, and distilled water up to a total sample volume of 20 µL.

Running an annealing-temperature gradient identified the optimal annealing-temperature, and the final cycling conditions were; initial denaturation at 98°C for 2.30 min, 39 cycles of 15 s denaturation at 98 °C, 30 s combined annealing and extension at 60 °C. All PCR reactions were performed in a Bio-Rad CFX96 Touch instrument (Bio-Rad Laboratories, Inc., Hercules, CA).

Assay performance

A custom ordered DNA-oligo (ThermoFischer, Hvidovre, DK) corresponding to a consensus of alignment for the amplicon of region A plus 10 bp 5'- and 3'-overhangs was used for generating a standard curve. Efficiency was assed by running 10-fold dilution of the plasmid (10^9-10^0 copies) and plotting the C_t-values versus log10 DNA copies thereby creating a standard-curve and exploring the assays dynamic range.

Specificity, sensitivity and reproducibility

Intra- and inter-assay reproducibility was assed by running multiple replicates of the same sample on one run and by running the same samples at 10 different runs, respectively. Sensitivity was assed by spiking known AMDV-negative spleen/serum-homogenates with a known amount of the DNA-oligo and running a standard-curve as described above. The previously implemented diagnostic endpoint PCR-assay (Jensen et al. 2011) was ran in parallel on the very same DNA-extractions in order to asses the analytical sensitivity. Specificity was assed by including non-template controls (ddH_2O) in all runs, and by using DNA extracted from known AMDV-negative mink in addition to DNA samples from mink infected with other related parvoviruses, e.g. distemper.

Diagnostic performance was assessed on clinical samples from 300 different animals confirmed to be positive for AMDV using serology (Dam-Tuxen et al. 2014). The diagnostic sensitivity was calculated as the number of true positive results, i.e. those identified as positive by both the conventional endpoint PCR and the real-time PCR, divided by the sum of number of true positives and false negative results. The diagnostic specificity was calculated as the number of true negative results, i.e. those identified as negative using for each of the conventional endpoint PCR and the real-time PCR, divided by the sum of number of true negatives and false positive results.

Data analysis

Sequencing raw-data generated by Sanger sequencing was imported into Geneious 7.1.5 (Kearse et al. 2012) where low quality bases in the ends and primer-sequences manually were removed. The forward and reverse strains were merged into a consensus to be used for further analysis. Alignments were visualized in Geneious 7.1.5 (Kearse et al. 2012). Raw data from real-time PCR runs were analysed in the Bio-Rad CFX Manager™ software (Bio-Rad Laboratories Inc., Hercules, CA).

Results

Selection of primers and probes

We developed a quantitative real-time hydrolysis PCR assay for detection of the three currently known circulating AMDV genotypes (here referred to as Saeby, Holstebro,

and Zealand). Primers generating amplicons in four genomic regions were evaluated individually using SYBR®-green chemistry and the two best regions were evaluated using TaqMan chemistry. The overall best primer-pair (table 1) was identified by the combined assessment of C_q-values, melting-curve analysis, and gel-electrophoresis (fig. 2) of PCR-fragments generated using SYBR®-green chemistry. The regions B, C, and D were excluded due to high C_q-values, unspecific amplification and the failure to detect the Zealand strain. To confirm PCR-amplification of the correct AMDV-genotypes the amplicons generated by primers A were Sanger sequenced and aligned to the available whole genome dataset used for primer-design and they were all in agreement (fig. 2).

Analytical performance of the real-time PCR assay

Assay efficiency was assed by running 10-fold dilution standard curve of a synthetic oligo representing a consensus sequence of the three genotypes. The dynamic range was 10^1-10^9 copies pr uL with a strong inverse linear relationship between the starting quantity and the C_q-value (fig. 3, panel A) reflecting accurate quantification over a large range of starting concentrations. The amplification efficiency was 106.2% and the coefficient of linear regression (R^2) was 0.996. The assay detected 10^1-10^2 copies of the custom made DNA oligo, and 10^x copies of DNA extracted from spleen. Assay efficiency was assed by running 10-fold dilution of the plasmid (10^9-10^0 copies) and plotting the C_t-values versus log10 DNA copies thereby creating a standard-curve and exploring the assays dynamic range.

The detection limits when running the assay on DNA extracted from clinical samples was 10^x copies for spleen and 10^x copies blood samples (no results yet), compared to the conventional PCR where the limit was 10^x copies.

Specificity was confirmed by no signal in the non-template controls (ddH_2O), and by running the assay on DNA samples extracted from known AMDV-negative mink and DNA samples from mink infected with other related parvoviruses (no results yet). Known AMDV positive samples from 300 different animals were tested in parallel using the real-time PCR assay developed here and the conventional end-point PCR (Jensen et al. 2011). X samples were in concordance, and the Y inconclusive samples

were removed from this part of the analysis as they in the clinical setting would have been investigated further or additional samples would have been requested prior to diagnosis. The diagnostic sensitivity was XX% and the specificity XX%.

Acknowledgements

Kopenhagen Diagnostics, Kopenhagen Fur, is thanked for supplying sample material and providing laboratory facilities. The laboratory technicians Carina Boegh Folsing (Kopenhagen Fur) and Sari Mia Dose (DTU Vet) are acknowledged for their valuable work in the laboratory.

References

Alexandersen, S., Bloom, M.E. & Perryman, S., 1988. Detailed transcription map of Aleutian mink disease parvovirus. *Journal of virology*, 62(10), pp.3684–94. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=253511&tool=pmcentrez&rendertype=abstract [Accessed November 4, 2014].

Anon, 2009. *Danish Executive Order 1447 of 15/12/2009*, Available at: https://www.retsinformation.dk/Forms/R0710.aspx?id=129366 [Accessed May 21, 2015].

Bloom, M.E. et al., 1988. Nucleotide sequence and genomic organization of Aleutian mink disease parvovirus (ADV): sequence comparisons between a nonpathogenic and a pathogenic strain of ADV. *Journal of virology*, 62(8), pp.2903–15. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=253728&tool=pmcentrez&rendertype=abstract.

Bloom, M.E., Race, R.E. & Wolfinbarger, J.B., 1980. Characterization of Aleutian disease virus as a parvovirus. *Journal of virology*, 35(3), pp.836–43. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=288877&tool=pmcentrez&rendertype=abstract [Accessed March 20, 2015].

Christensen, L.S. et al., 2011. Diversity and stability of Aleutian mink disease virus during bottleneck transitions resulting from eradication in domestic mink in Denmark. *Veterinary microbiology*, 149(1–2), pp.64–71. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21112164 [Accessed May 21, 2014].

229 Dam-Tuxen, R. et al., 2014. Diagnosing Aleutian mink disease infection by a new fully
230 automated ELISA or by counter current immunoelectrophoresis: A comparison of
231 sensitivity and specificity. *Journal of Virological Methods*, 199, pp.53–60.

232 Decaro, Nicola et al., 2012. 12. Parvovirus infections. In D. Gavier-Widén, J. P. Duff, & A.
233 Meredith, eds. *Infectious Diseases of Wild Mammals and Birds in Europe*. Oxford, UK:
234 Wiley-Blackwell, pp. 181–285.

235 Fields, B.N., Knipe, D.M. & Howley, P.M., 2007. Fields Virology, 5th Edition. *Fields*
236 *Virology*, 2, p.3177. Available at:
237 http://www.loc.gov/catdir/toc/ecip072/2006032230.html.

238 Hagberg, E.E., Larsen, L.E., Krarup, A., Pedersen, A.G., 2017. *Genetic analysis of the entire*
239 *genome of Aleutian Mink Disease Virus determines its evolutionary rate and confirms*
240 *bottleneck due to control program*,

241 Hagberg, E.E. et al., 2016. A fast and robust method for whole genome sequencing of the
242 Aleutian Mink Disease Virus (AMDV) genome. *Journal of Virological Methods*.
243 Available at: http://www.sciencedirect.com/science/article/pii/S0166093415300343.

244 Hagberg, E.E., 2017. *Molecular epidemiology of AMDV*. Technical University of Denmark.

245 Jensen, T.H. et al., 2011. Implementation and validation of a sensitive PCR detection method
246 in the eradication campaign against Aleutian mink disease virus. *Journal of virological*
247 *methods*, 171(1), pp.81–5. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20951744
248 [Accessed May 21, 2014].

249 Jensen, T.H., Chriél, M. & Hansen, M.S., 2015. Progression of experimental chronic Aleutian
250 mink disease virus infection. *Acta Veterinaria Scandinavica*, 58(1), p.35. Available at:
251 http://actavetscand.biomedcentral.com/articles/10.1186/s13028-016-0214-7.

252 Kearse, M. et al., 2012. Geneious Basic: an integrated and extendable desktop software
253 platform for the organization and analysis of sequence data. *Bioinformatics (Oxford,*
254 *England)*, 28(12), pp.1647–9. Available at:
255 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3371832&tool=pmcentrez&
256 rendertype=abstract [Accessed July 10, 2014].

257 Ryt-Hansen, P., Hagberg, E.E., et al., 2017. *Global Phylogenetic analysis of contemporary*
258 *Aleutian Mink Disease Viruses (AMDVs)*,

259 Ryt-Hansen, P., Hjulsager, C.K., et al., 2017. *Outbreak investigation of Aleutian Mink*
260 *Disease Virus (AMDV) using partial NS1 gene sequencing*,

261 Ye, J. et al., 2012. Primer-BLAST: a tool to design target-specific primers for polymerase

chain reaction. *BMC bioinformatics*, 13, p.134. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3412702&tool=pmcentrez&rendertype=abstract [Accessed November 8, 2014].

Tables and figures

TABLE 1. Primer sequences and expected amplicon sizes.

Primer	Position (NC001662)	Primer sequence (5'-3')	size (bp)
A3F	2836-3856	ACTTAACTGCGTCGTTACAGG	21
A2R	2908-2927	AACAAAGCCCAGTGTTTCCC	20
A2P-f	2874-2900	CCGGGGGGCACTGGAAAAACCTTG	24
amplicon	2836-2927		91
oligo	2826-2937		121

Table 1. The primer sequences designed in the present study and the sizes of the expected amplicons for the applicable combinations. Forward primers are indicated F, reverse primers R. All primers were designed using the Primer 3 software and a whole genome alignment of three known AMDV strains (Hagberg, E.E., Larsen, L.E., Krarup, A., Pedersen 2017).

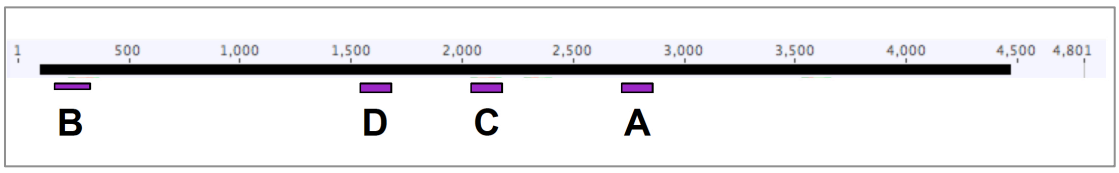


Figure 1. Location of tested amplicons. Schematic illustration of the AMDV-genome and location of the four genomic regions targeted during initial primer-screening. Figure created using Geneious v.7.1.5 (Kearse et al. 2012).

	NC_001662	15_DK_Sa…	Batch_A01…	15_DK_Mi…	Lot_nr_A0…	15_DK_Ze…	Lot_nr_A0…
NC_001662		93.1%	93.7%	86.5%	86.8%	87.1%	87.1%
15_DK_Saeby_field-st…	93.1%		99.3%	84.8%	84.8%	86.1%	86.1%
Batch_A012-026_2_S…	93.7%	99.3%		85.5%	85.5%	86.1%	86.1%
15_DK_Mid_Jutland_fi…	86.5%	84.8%	85.5%		99.3%	91.7%	91.7%
Lot_nr_A012-033_1_…	86.8%	84.8%	85.5%	99.3%		91.1%	91.1%
15_DK_Zealand_field-…	87.1%	86.1%	86.1%	91.7%	91.1%		100%
Lot_nr_A012-032_1_…	87.1%	86.1%	86.1%	91.7%	91.1%	100%	

Figure 2. Confirmation of amplification by the PCR-primers. The distance-matrix created in Geneious v.7.1.5 (Kearse et al. 2012) shows the percentage agreement between a representative sequences from each AMDV genotype in the alignment used for primer-design and the corresponding PCR-amplicons generated with the PCR-primers designed in this study. DNA alignment on behalf of Emma Hagberg (Hagberg, E.E., Larsen, L.E., Krarup, A., Pedersen 2017; Hagberg 2017).

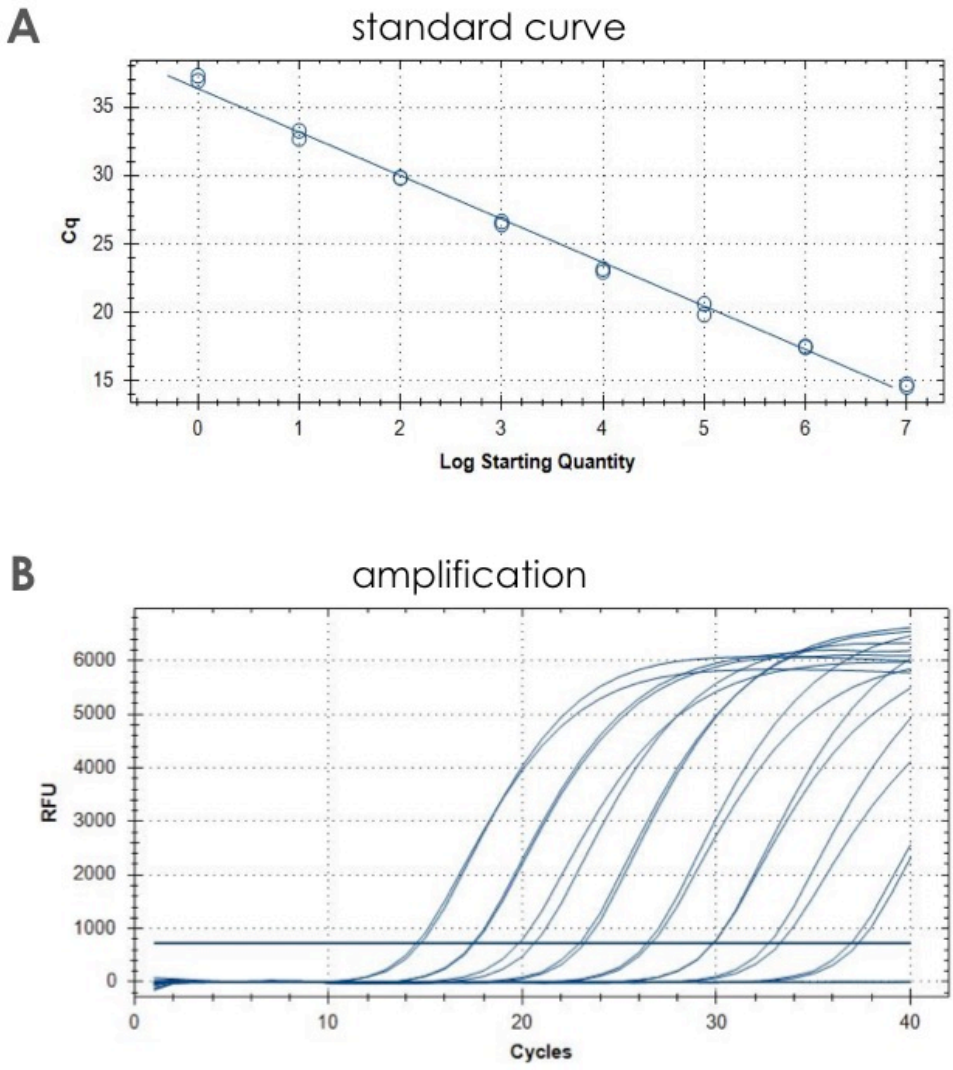


Figure 3. Standard curve and amplification plot. Panel A shows the linear relationship between the log starting quantity and the C_q value. The circles represent samples. Panel B shows the RFU for the amplification curves for the triplicates in each dilution in the series. The horizontal line indicates the detection threshold. Plots generated in Bio-Rad CFX Manager™ software (Bio-Rad Laboratories Inc., Hercules, CA).

Chapter 5

DISCUSSION AND CONCLUSIONS

5. DISCUSSION AND CONCLUSIONS

5.1. DISCUSSION

Aleutian Mink Disease virus (AMDV) is globally the most important pathogen related to mink farming. In Denmark AMDV has been monitored since 1999 by a national control program, which is based on serological screening of all animals, and encourages infected farms to depopulate. In terms of genomic surveillance, there has historically been no consensus between the mink farming countries about which genomic region of the virus to analyse in this regard, and most previous studies have been based on partial (Jensen et al. 2011) or entire genes (Sang et al. 2012; Leimann et al. 2015; Knuuttila et al. 2015; Oie et al. 1996), or on pure epidemiological data (Christensen et al. 2011). Thus, when initiating this project, little was known about AMDV's total genomic diversity and how the virus was spread between farms. The overall aim was to investigate if recent advances such as next generation sequencing and more advanced phylogenetic analyses of full-length isolates could improve our understanding of the total genomic diversity and evolution of AMDV, and to evaluate if this knowledge could contribute to elucidate AMDV transmission between farms and improve molecular diagnostics.

The fast and robust method for whole-genome sequencing of the AMDV genome published as Manuscript 1, enabled the sequencing of a large number of viral isolates and provided the necessary foundation for the remaining analyses in this thesis. To illustrate the superiority of using whole genome sequences compared to the in Denmark traditionally used partial NS1 gene sequence for elucidating transmission patterns between farms, a proof-of-concept study was performed on a limited set of AMDV samples. Manuscript 2 is the result of this work, clearly illustrating that the phylogenies based on partial NS1 gene sequencing were uninformative, and could not be used for determining transmission pathways, not even in the light of supporting epidemiological data (Hagberg et al. 2017). The explanation for this was the high sequence homology, which made the phylogenetic inference impossible as the leaves were branching out from the same internal nodes. When repeating the analysis using whole genome sequences from the same isolates, the additional information contained in these longer sequences almost completely resolved the phylogeny. Incorporating sampling-dates into the phylogenetic analysis improved WGS-tree resolution further, in addition to providing an age-estimate for the MRCA of the infected farms, and thus allowed us to confirm the epidemiological hypothesis about the direction of spread (Hagberg et al. 2017). The viral isolates in Manuscript 2 where all of the Saeby type (Christensen et al. 2011), and it was clear that despite using the longer whole genome sequences, the low pairwise differences emphasise the

inclusion of epidemiological meta-data such as AMDV prevalence (extrapolated from the annual serological testing) and farm locations, if the phylogenetic inferences are to be used for supporting directions of spread.

Manuscript 3 adopts the methodologies and workflows from Manuscript 1 and 2, and is the to-date most comprehensive full-length phylogenetic analysis of AMDV, composed of more than 200 field isolates. It contributes to our understanding of the AMDV genome and can facilitate future diagnostics of the virus. For example, the real-time PCR assay described in Manuscript 4 was developed based on the additional nucleotide information gained from sequencing entire AMDV genomes of different strains.

How to use phylogenetic trees in relation to an AMDV outbreak?

In relation to viral outbreaks or epidemics it is common practise to use a consensus sequence for each sample for downstream analysis (Gilchrist et al. 2015), as was done throughout this thesis. It should however be kept in mind that such consensus sequences represents an "average DNA sequence" of the virus of interest in that particular sample, and thus might not represent an *in vivo* existing viral species. This has implications for the interpretation of pairwise comparisons and phylogenetic trees, e.g. when trying to identify the source of an outbreak as there almost certainly will not be a direct match between the samples. One could speculate that the majority species in an individual at the point of sampling would have a higher probability of being captured by the PCR, and if substantially evolved from an earlier minority ancestor, the pursuit for a perfectly MRCA would be challenging. On the contrary, if the PCR-primers were designed in well-conserved genomic regions, each position in the viral consensus sequence will represent the most dominant base in the underlying population of reads, and thus also the population. When the sequence diversity is low, as in the case of AMDV, the consensus will likely be a good proxy for each sample.

One approach for identifying outbreak clusters is to determine the non-epidemic genotype most closely related to the epidemic genotype. However, this heavily depends on the collected data (Kühnert et al. 2011), and if one fails to identify a strain sufficiently related to the epidemic strain the phylogenetic tree cannot provide the origin. It is easier to establish the direction of spread if samples are collected close enough in time to the transmission event, as a subset of source sequences will be more closely related to the recipient sequences than all source sequences are to each other (Metzker et al. 2002). However, the window where this is possible is affected by the substitution rate of the viral genome and immune-selection pressure from the host. Despite the low sequence diversity in AMDV, its in comparison to other DNA viruses relatively high clock-rate (Hagberg, E.E., Larsen, L.E., Krarup,

A., Pedersen 2017) probably helped to resolve the phylogenies. Altogether, this underlines that denser and more regular sampling as well as full-length sequencing of AMDV isolates is required to build a continuously updated library of AMDV sequences if one is to use genetic molecular analyses for surveillance.

Despite improving tree resolution, it has been shown that in many cases, whole genome-based phylogenetics cannot stand alone, but needs to be supplemented with solid epidemiological data (Leekitcharoenphon et al. 2014; Metzker et al. 2002). The disease status and prevalence on the farms is a useful indicator for the direction of spread, as a farm is more likely to be the source if it had higher prevalence at the time of time of spread (Ypma et al. 2013; Hagberg et al. 2017). By assuming that sequences evolve according to a clock model, and by using tip sampling-dates to calibrate the phylogeny, the relationships between the sequences can be visualised on a time-scale, and the age of the MRCA can be estimated and correlated with epidemiological information such as disease status of the farm. When performing dated analyses, one makes a strong assumption about the dates being correct, and should be aware that they can, and will, influence the phylogeny. This assumption can however be relaxed by putting a prior distribution on the dates, and if the overall tree topology remains the same as in an undated phylogeny based on the same data, it is an indicator that the tip-dates are consistent with the evolutionary information contained within the sequences. Using tip-dates on AMDV samples increased the tree resolution without changing the overall tree topology, and allowed for the age of the MRCA of a group of case-farms to be estimated with presumed high accuracy (Manuscript 2). The sequences from Manuscript 2 were subsequently reanalysed in Manuscript 3, and despite being surrounded by a context of an additional 150 viral isolates; the relationships between the three case-farms remained the same as in Manuscript 2. These findings support the robustness of the methodology, and suggest the potential of WGS and phylogenetics for AMDV surveillance.

Does it have to be so complicated?

Bayesian phylogenetic methods quantify and describe the degree of belief in the results. However, no results are better than the specified model and input data, and thus applying a realistic model and carefully setting the priors is important. Overall, it is advised to choose a model that is complicated enough to describe the data, while still allowing for identification of unknown parameters of interest (Huelsenbeck & Rannala 2004). People new to the field sometimes consider phylogenetics as a "black box". However, there are some helpful assumptions to keep in mind, and this section will try to address some of them.

How to select nucleotide a substitution model? For most data analysed in this thesis, the model-testing results indicated either a GTR or HKY model, which could be a sign of different parts of the genome exhibiting different models of evolution. Mainly the HKY-model was applied, as it often describes viral dynamics well, and the choice between GTR and HKY should not matter too much if site variation is modelled using a Gamma-shape distribution with an L-shaped prior simultaneously taking into account invariant sites (Drummond et al. 2007).

The AMDV samples in this study represent intra-species datasets with low variation between branches. This type of biology has previously been suggested best to be described with a strict molecular clock and a coalescent population model (Brown & Yang 2011; Drummond et al. 2007), especially if the data is a subsample of the population, as was the case in this thesis. If the data is sufficiently informative and the dating is accurate, the priors set on mutation rate and clock rate will not have high impact on the results (Drummond & Bouckaert 2015), something which easily can be tested by setting different priors on the rates and confirm they converge to similar values, as verified in Manuscript 3.

In regards to molecular clock and tree priors, presumably neither a constant or exponential tree population model, nor a strict molecular clock, are perfect for describing the dynamics in a viral population. But to successfully infer parameters using the more complex birth-death models, additional knowledge about population parameters would be required. Kopenhagen Fur keeps records of the serological testing in Denmark, and to thoroughly go through these records and create an overview of the AMDV status for each year, taking into account the fraction of sequenced samples would be a useful direction for a future study.

Partial gene or WGS for AMDV diagnostics?

During the 2015-2016 AMDV outbreak, two new strains were detected on Danish farms using partial NS1 gene sequencing (Ryt-Hansen, Hjulsager, et al. 2017). These strains were most closely related to those from other mink producing countries (Ryt-Hansen, Hagberg, et al. 2017), supporting the epidemiological hypothesis that the disease was imported to Denmark. The authors reported that in contrast to the Saeby strain, the two new strains exhibited high genetic diversity and were distinguishable even at farm level using phylogenies constructed from the partial NS1 gene (Ryt-Hansen, Hagberg, et al. 2017). It was however impossible to elucidate transmission patterns based on these partial NS1 gene data, and it was concluded that partial NS1 gene sequencing could solely be used for differentiating between major viral clusters. Thus, supporting the findings from Manuscript

2, that the additional genetic information such as that contained in WGS is needed if the AMDV genome is to be used for suggesting transmission patterns between farms (Hagberg et al. 2017).

Intuitively it might seem logical to try to resolve difficult phylogenies by including additional sequences in the analysis (Rosenberg & Kumar 2003; Heath et al. 2008). However, if the sequence itself does not contain enough information, this approach will be unsuccessful (Manuscript 2). Instead the solution is to increase the number of informative sites. The effect of increasing sequence length is illustrated in Manuscript 2, where WGS and partial NS1 gene phylogenies were created from the very same data – an alignment of full length sequences from which the sections corresponding to the previously used NS1 region was cut out. The superiority of WGS was clear, however not surprising, as using short sequences to reconstruct the phylogenetic tree of closely related species decrease the probability of that tree being equal to the species tree (Pamilo & Nei 1988). Others have confirmed this lack of resolution when using partial (Jensen et al. 2011) or single genes (Sang et al. 2012; Leimann et al. 2015; Knuuttila et al. 2015; Oie et al. 1996) for constructing AMDV phylogenies. Furthermore, low phylogenetic resolution has been demonstrated even between different canine parvoviruses when using full length VP2-sequences for the reconstruction (Allison et al. 2013), showing that the VP2-gene is relatively well conserved. Canuti et al. (2016) reported "well supported incongruities" depending on which part of the AMDV genome, and even which part of the genes, was used for the phylogenetic inference. The sliding-window approach applied both in Manuscript 2 and Manuscript 3 showed there were no other genomic regions or limited number of informative sites that could be combined to provide equally good phylogenetic resolution as WGS. Unfortunately, due to external factors such as time constraints, only a few of the foreign strains from the Ryt-Hansen et al. (2017) study where whole-genome sequenced and included in the present PhD-thesis (Manuscript 3). However, it would be interesting to sequence an additional of these outbreak associated strains in full-length and to re-estimate their phylogeny, as the sliding-window analyses from this thesis illustrated how the tree topology changes across the genome (Manuscript 3).

While it is clear that WGS provide useful phylogenetic resolution in regards to tracing AMDV transmission, it is important to take diagnostic aim and practicalities into account. If the aim is sensitive detection of virus, then WGS is not the best choice due to it being a longer fragment and hence more difficult to amplify. Furthermore, the timelines of the sampling is important, and NGS is time-consuming, expensive, and requires an up-to-date database in order to fully utilise its potential in regards to outbreak investigation (Gilchrist et al. 2015). The WGS approach presented here relies on a slower and more expensive PCR-amplification step, and the per-sample price gets high when the

NGS step is performed on small batches, which would be the likely scenario during a 'normal' small-scale outbreak in Denmark. Additionally, the NGS data-analysis requires not only substantial computing power, but also specialised expertise throughout the data processing and analysis steps.

A more economical and feasible approach could be to use a faster, cheaper, and more sensitive protocol, such as the partial NS1 PCR (Jensen et al. 2011), for continuous surveillance and for monitoring deviations from the expected strains. Alternatively, a multiplex real-time PCR with the ability to discriminate between the most well characterised AMDV strains could be used for the same purpose. Samples could then be pooled on a routinely basis, and sequenced in full-length using NGS when there are enough to fill up a sequencing chip. On the contrary, the prices for WGS are expected to decrease with time, and cheaper platforms may also be developed which will ease the use of WGS in routine diagnostic settings. In conclusion, in a diagnostic setting it is useful to a priori define the purpose with the DNA based diagnostic approach, e.g. sensitive detection, viral typing, or elucidating transmission patterns, as this will influence the choice of strategy.

The rate of molecular evolution

Accurate knowledge about a pathogens evolutionary rate is important for tracing its transmission as this allows phylogenies to be converted to time-scales, and thus specific time-points for transmission events can be estimated. Most research on viral evolutionary rates has been performed on an individual gene level, which is not necessarily the best proxy for an entire organism, and thus a genome wide approach has been suggested crucial to fully understand viral evolutionary dynamics (Duffy et al. 2008). The data presented in this thesis are the first to report AMDV evolutionary rates based on analyses of whole genome sequences isolated from *in vivo* infectious strains, and suggest a rate in the magnitude of 10^{-4} subs/site/year (Hagberg, E.E., Larsen, L.E., Krarup, A., Pedersen 2017). This corresponds to approximately 0.5 mutations per genome per year, which is higher than many other DNA viruses, but in line with other parvoviruses (table 1). Since substitution rates can be overestimated when data is collected over a short time-span (Duffy et al. 2008), as often is the case during an active outbreak, it is important to have robust estimates for the viral evolutionary rate a priori, in order to predict how fast the virus changes and spreads, and for estimating the origin of the outbreak. The samples in Manuscript 2 (Hagberg et al. 2017) were collected during the same year, and thus the time-stamped phylogeny from that study was primarily useful for determining relative, but not specific time-points for transmission. However, the MRCA estimate for case farms was supported by the epidemiological data, as it was biologically realistic that it was younger than one year due to the stamping out policy in Denmark. A similar scenario was observed in Manuscript 3, where the

Saeby strain had a fivefold lower clock-rate compared to the strains related to the 2015/2016 outbreaks. However, also in this case a plausible explanation was that the outbreak sequences were collected during a shorter time-span (2015-2016) impacting the rate estimates (Duffy et al. 2008). But the higher rate in the outbreak cluster could also be biologically "real" and due to different population dynamics, such as the outbreak strains being more virulent, or due to a more naïve animal population since the outbreaks mainly involved farms in a previously AMDV free region (Ryt-Hansen, Hjulsager, et al. 2017).

Another interesting implication of molecular evolutionary rates is how they can differ between genes, and thus can impact the phylogeny and dating of the MRCA's depending on which parts of the genomes are used for the inference. Canuti et al. (2016) discuss that the differences between the NS and VP parts could be due to different evolutionary histories and rates of the two regions, i.e. recombination. In the present study the substitution rates did not substantially differ between the NS1 and VP2 genes (Manuscript 3), and were in line with a previous findings, such as an analysis of a related parvovirus, porcine parvovirus (PPV), where the VP1 gene had a rate of 10^{-4} subs/site/year (Streck et al. 2011).

Host factors and selection pressure

The genetic data in Manuscript 3 showed that the AMDV population could be divided into two major clusters separated with a deep root. These two clusters exhibited slightly differing dynamics, where overall, the strains most closely related to the in Denmark familiar Saeby strain, had less genomic variation and seemed to be subject to less evolutionary pressure compared to the outbreak related strains. Furthermore, the analyses of the Saeby strain were consistent with the efficiency of the Danish control programme, i.e. the effective population size decreased a few years after the programme's implementation in 1999. The skyline plot has been applied in other viruses both to describe their past and to relate bottlenecks to interventions such as vaccination or a control programme (Strimmer & Pybus 2001). When the expected effective population size is low, for example due to transmission bottlenecks and pelting down each season (Nelson & Hughes 2015), there would every year be a new naïve host population. This could, in addition to the presumably low genetic diversity in farmed mink and their immune systems (e.g. due to years of intensive breeding), favour efficient viral replication and spread within the farm. The deep root separating the clusters, however also suggested there was a large number of unsampled hosts, and that the sampled strains originated from different ancestors separated for a long period of time.

The Danish AMDV eradication effort, i.e. years of reductive pressure through the removal of infected individuals from the population, was reflected by a decrease in effective population size just a few years after the control programme was implemented in 1999. This bottleneck resulted in very high sequence homology within the Saeby strain (Hagberg, E.E., Larsen, L.E., Krarup, A., Pedersen 2017). The strains that managed to evade the eradication efforts, e.g. through survival in the soil or the interior of the farms, could be the strains best adapted to resist environmental conditions or best at avoiding the mink's immune systems. This would in particular apply to the strains in the endemic area of Northern Jutland, where the AMDV prevalence on the farms and thus also the eradication pressure is higher. The investigation of selective pressure was limited to the genes encoding the non-structural (NS1) and the viral capsid (VP2) because of their presumed importance for viral replication and entry, and due to practical considerations such as most AMDV genes being encoded by alternative splicing (Alexandersen et al. 1988; Qiu et al. 2006). In agreement with previous studies (Christensen et al. 2011; Knuuttila et al. 2009; Canuti et al. 2016), the findings from Manuscript 3 suggested that the genome was mainly subject to a negative or purifying selection pressure. This supports the hypothesis that most mutations would have a negative impact on the fitness of the virus in that particular population, and that the virus is highly optimised in terms of critical sites for survival and avoidance of the host (Fields et al. 2007). Additionally, negative selection might contribute to keep the substitution rate in the current range, as a positive selection within the host (i.e. avoiding the hosts response) could lead to an increased substitution rate (Duffy et al. 2008).

It was however challenging to assess whether the differences in selective pressure *between* the clusters was due to a *real* difference, or if it reflected an impaired ability to *detect* the difference, due to the low genomic diversity within the Saeby cluster (Streck et al. 2011; Kryazhimskiy & Plotkin 2008). A similar challenge was encountered in regards to recombination detection. Results from studies in other parvoviruses are somewhat conflicting, some have suggested that recombination might have an important role (Canuti et al. 2016; Shackelton et al. 2007), while others have not been able to detect it in e.g. Porcine Parvovirus (PPV) (Streck et al. 2011). We were not able to detect recombination in the data presented here, which perhaps reflect that recombination detection, just like selection pressure detection, has low power when the genomic diversity decreases (Posada et al. 2002). Other factors which might influence the results and their interpretation, is the use of NGS and consensus sequences, as these sequences might not reflect *in vivo* viable strains (further discussed below), and that the sequences in the more diverse outbreak-cluster mainly originated from a single outbreak while the Saeby cluster represented several years of sampling from a persistent population.

A reoccurring question is the one about differing pathogenicity between the AMDV strains. Just to mention a few, the VP2 amino acids 115 and 491 have previously been linked to pathogenicity and were found to differ between the strains investigated in Manuscript 3. The strains related to the Danish 2015/2016 outbreak had the same amino acid as the pathogenic AMDV-Utah at these positions, while the Saeby strain and the non-pathogenic AMDV-G had another. But whether those changes can be directly linked to pathogenicity is difficult to conclude, especially since the *perceived* higher pathogenicity in these outbreak strains most likely was influenced by other factors, such as the animals immune system, genetics, and farm management.

Farm management factors

The control of AMDV impose a major challenge on the farmers, as it has several properties for successful survival and evolutionary progression: e.g. it is shed from chronically infected hosts for a long period of time, it seems to elude the hosts immune defence, it survives in the external environment, and is capable of vertical transmission (Maclachlan et al. 2011). AMDV stability in the environment enables efficient host-to-host transmission within the farm facilitated by management factors such as handling and movement of animals both *within* and *between* farms, trade, and movement of personnel and materiel. Practical operation factors of a farm that can impede biosafety and should be considered carefully in the daily routines are e.g. the open barns allowing for plentiful ventilation, possible wildlife access, external feed-supply transportation routes, and the proximity between the farms (illustrated in Manuscript 2, fig. 1). For example, in a study of avian influenza the wind direction at the date of transmission was correlated to the between-farm viral spread of the virus (Ypma et al. 2013). Taking into account the close proximity between the mink farms, especially in Northern Jutland, possible influence through the wind or by vectors seems plausible (Manuscript 2), but has not been thoroughly investigated. The importance of a wildlife reservoir of AMDV has been discussed previously (Ryt-Hansen, Hjulsager, et al. 2017; Farid 2013; Jensen et al. 2011), but without revealing a clear link to the AMDV strains isolated from farmed mink. This could however be due to a potentially large number of unsampled hosts in the wild fauna, that if sampled could split the existing clusters.

Sample preparation and sequencing technology

In addition to the phylogenetic reconstruction process and the sampling, the input (i.e. the samples) requires consideration. The use of specific PCR amplification was important for future field applications, to avoid interfering host DNA, to generate sufficient amounts of double stranded DNA for sequencing library preparation, and to overcome the labour intensive steps of amplifying the

ADMV-genome multiple fragments as previously done by others (Li et al. 2012; Canuti et al. 2016). To some extent, sequencing of PCR amplified DNA might influence the ability to capture the full intra-individual viral variation, which could be larger due to e.g. recombination, quasispecies, or simultaneous infection with more than one strain. However, as discussed above, if the PCR-primers are designed in well-conserved genomic regions, the risk of losing potential genomic diversity should be small, and the consensus sequence will be a good proxy. Furthermore, it should be remembered that the analyses in this study were performed using consensus sequences generated by NGS, which equals to analyse an "average viral strain" from the sample/animal it was isolated from, and thus, might not represent *in vivo* viable strains. This can introduce artefacts such as breaks in open reading frames, as seen in Manuscript 3, but should not substantially influence the phylogenies when nucleotide data is used for the inferences. However, codon- and protein-based analyses can become difficult. Nonetheless, PCR-amplification followed by NGS, and phylogenetic analysis of consensus sequences has successfully been applied to determine transmission patterns for many other pathogens (Gire et al. 2014; Hagberg et al. 2017; Leekitcharoenphon et al. 2014; Escobar-Gutiérrez et al. 2012).

The first two thirds of the AMDV genome were easier to amplify compared to the last third of the genome, i.e. there was a higher PCR success-rate with the so-called "fragment A". Possible explanations could be that the 5'-primer annealed better than the 3'-primer. However, during development of the long-range PCR assay these primers amplified well, both the cell-culture adapted AMDV-G and the highly virulent AMDV-Utah, and thus were considered able to capture a broad span of genotypes (Hagberg et al. 2016). The overall genomic diversity, both at a nucleotide and amino acid level, were higher in the first 2/3 of the genome (Hagberg, E.E., Larsen, L.E., Krarup, A., Pedersen 2017; Hagberg et al. 2016), which is consistent with previous findings suggesting the AMDV NS-gene is more variable than encountered in other parvoviruses (Gottschalck et al. 1994). Therefore, more plausible explanations for the lower primer-binding hit rate in the 3'-end could be the presence of secondary structures in the second ORF, or even sample degradation (Wohl et al. 2016).

A caveat with the chosen sequencing technology, the Ion Torrent PGM, is that it is known to struggle with accurately registering the number of consecutive bases in homopolymeric and GC-rich regions (Quail et al. 2012; Hagberg et al. 2016). To verify the accuracy of the Ion Torrent PGM, the NGS consensus sequences were verified by Sanger-sequencing of the known homopolymeric GC-rich region (AMDV-G: nt 2470-2520). In order to further verify the NGS approach, the sequences from Manuscript 3 (generated by NGS), were aligned to and were in agreement with their corresponding

Sanger generated partial NS1 sequences from another contemporary project at DTU Veterinary (Pia Ryt-Hansen et al.), suggesting the Ion Torrent produced reliable data.

It could have been useful initially to benchmark the Ion Torrent PGM with another sequencing technology. But at that stage of the project, the decision to use the Ion Torrent was most straightforward due its competitive price and the availability (both of the machine and technical expertise) at the DTU Core facility (DMAC). For example, the MinION by Oxford Nanopore could have been an interesting alternative to test, as it allows for real-time sequencing of a single DNA molecule. However this would still require an additional pre-sequencing step in order to overcome interfering host-DNA.

Future perspectives/directions

Routinely implemented whole genome-based surveillance would enable us to be on the forefront in case a new outbreak emerges. This approach would require a shift in paradigm, from reacting and typing individual cases, to proactively acting upfront and performing whole genome based monitoring on a regular basis. The unambiguity in determining the origin and links between outbreaks in different countries, especially when there is incomplete sampling, is a challenge for the surveillance of many pathogens (Gilchrist et al. 2015) including AMDV (Ryt-Hansen, Hagberg, et al. 2017). This problem could be addressed, e.g. by the establishment of an international WGS database, which could enable a real-time overview of the global AMDV genomic diversity, and thus benefit all fur producing countries regardless of prior genomic surveillance strategy (if any at all). Despite that the phylogenetic methodology seems fairly robust, accurate annotation of sequences, hereunder registration of meta-data such as location and sampling-dates, is crucial for the validity of such a surveillance database, as inferences could be skewed by incorporating inaccurate meta-data due the low sequence diversity reported in this thesis. Furthermore, timely sequencing and data analysis is essential if such a database is to add value, and thus requires dedicated resources such as staff and computing facilities to manage, analyse, and interpret the data and to communicate the results.

Another challenge with AMDV whole genome based protocols, is that, like the partial gene approach, the animals are required to be euthanized prior to sampling, which is not always feasible. The protocol developed during this thesis was intended for blood samples, however the often-low viral concentrations in blood made this approach unreliable. On the other hand, genomic surveillance of blood samples might be sufficient for farm-level surveillance, especially in heavily infected farms (i.e. with high disease prevalence) as there at most given time-points will be animals at different stages of

infection, and therefore could be captured by the presumably less sensitive blood sample protocol. This however, would require a larger investment in terms of personnel to process the higher number of samples, and therefore might not be feasible.

In regards to *in vivo* isolated field strains, the optimal scenario would be to thoroughly assess intra-farm variation before making conclusions about the inter-farm variation. However, due to practical considerations related to the Danish regulations, i.e. the farmers stamp out the population if an infection is discovered, it is hardly possible to get more than one or two animals from each individual farm. Therefore the overall pragmatic approach was to use the material that was accessible. Another aspect to address could be to map the distribution of AMDV within the animals, within the organs, and within the farms, as a better understanding of these aspects could aid in our understanding of the disease and to direct and rationalise sampling. For example, performing more extensive sampling of e.g. 10 animals per farm in 10 farms to allow for systematic investigation of the within- and between-farm diversity. Such experiments would be very interesting, but somewhat hard to conduct since the Danish authorities control the disease, and because infected animals are usually removed directly upon detection to limit the viral load on the farm.

5.2. CONCLUSIONS AND ACTIONABLE SUGGESTIONS

- The work presented in this thesis contributes to the molecular characterisation of AMDV and enables us to better understand its evolution.
- Phylogenies based on whole genome sequences accompanied by meta-data are required for elucidating transmission patterns between farms.
- PCR is detection of DNA, not infectious virus, while serological methods such as ELISA and CIEP detect the antibody response towards the virus. Thus PCR followed by WGS is a set of tools to supplement the existing diagnostic methods for AMDV, and if applied proactively, could improve our understanding of the pathogen and its transmission, as well as the diagnostic program carried out by Kopenhagen Fur.
- Define a clear purpose with the DNA based diagnostics and choose a strategy accordingly.
- Develop a multiplex real-time PCR assay to distinguish between AMDV strains.
- Systematically revisit historical AMDV records and generate sampling-proportions, and keep the database updated when additional samples are sequenced.
- Two-step surveillance programme:
  1) Frequent and continuous sub-typing using multiplex real-time PCR or partial NS1 gene sequencing.
  2) Pool and perform NGS on whole genomes to create database.

REFERENCES

6. REFERENCES

Aasted, B., 1980. Purification and characterization of Aleutian disease virus. *Acta pathologica et microbiologica Scandinavica. Section B, Microbiology*, 88(6), pp.323–8. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1887888&tool=pmcentrez&rendertype=abstract.

Alexandersen, S., Bloom, M.E. & Perryman, S., 1988. Detailed transcription map of Aleutian mink disease parvovirus. *Journal of virology*, 62(10), pp.3684–94. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=253511&tool=pmcentrez&rendertype=abstract [Accessed November 4, 2014].

Allison, A.B. et al., 2013. Frequent Cross-Species Transmission of Parvoviruses among Diverse Carnivore Hosts. *Journal of Virology*, 87(4), pp.2342–2347. Available at: http://jvi.asm.org/cgi/doi/10.1128/JVI.02428-12 [Accessed September 9, 2016].

Andrews, S., 2010. FastQC: a quality control tool for high throughput sequence data. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Anon, CLC Genomic Workbench. Available at: https://www.qiagenbioinformatics.com/ [Accessed September 5, 2015a].

Anon, 2009. *Danish Executive Order 1447 of 15/12/2009*, Available at: https://www.retsinformation.dk/Forms/R0710.aspx?id=129366 [Accessed May 21, 2015].

Anon, ICTV Virus Taxonomy 2015. Available at: http://www.ictvonline.org/virustaxonomy.asp [Accessed September 9, 2016b].

Berns, K.I., 1990. Parvovirus replication. *Microbiological reviews*, 54(3), pp.316–29. Available at: http://www.ncbi.nlm.nih.gov/pubmed/2215424 [Accessed September 9, 2016].

Bloom, M.E. et al., 1998. Construction of pathogenic molecular clones of Aleutian mink disease parvovirus that replicate both in vivo and in vitro. *Virology*, 251(2), pp.288–96. Available at: http://www.sciencedirect.com/science/article/pii/S0042682298994260 [Accessed May 13, 2015].

Bloom, M.E., Kaaden, O.R., et al., 1988. Molecular comparisons of in vivo- and in vitro-derived strains of Aleutian disease of mink parvovirus. *Journal of virology*, 62(1), pp.132–8. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=250511&tool=pmcentrez&rendertype=abstract.

Bloom, M.E., Alexandersen, S., et al., 1988. Nucleotide sequence and genomic organization of Aleutian mink disease parvovirus (ADV): sequence comparisons between a nonpathogenic and a pathogenic strain of ADV. *Journal of virology*, 62(8), pp.2903–15. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=253728&tool=pmcentrez&rendertype=abstract.

Bloom, M.E. et al., 1990. Nucleotide sequence of the 5'-terminal palindrome of Aleutian mink disease parvovirus and construction of an infectious molecular clone. *Journal of virology*, 64(7), pp.3551–6. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=249630&tool=pmcentrez&rendertype=abstract [Accessed May 27, 2014].

Bloom, M.E., Race, R.E. & Wolfinbarger, J.B., 1980. Characterization of Aleutian disease virus as a parvovirus. *Journal of virology*, 35(3), pp.836–43. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=288877&tool=pmcentrez&rendertype=abstract [Accessed March 20, 2015].

Bloom, M.E., Race, R.E. & Wolfinbarger, J.B., 1982. Identification of a nonvirion protein of Aleutian disease virus: mink with Aleutian disease have antibody to both virion and nonvirion proteins. *Journal of virology*, 43(2), pp.608–16. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=256163&tool=pmcentrez&rendertype=abstract.

Bouckaert, R. et al., 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS computational biology*, 10(4), p.e1003537. Available at: http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003537 [Accessed July 11, 2014].

Bouckaert, R., 2015. bModelTest: Bayesian site model selection for nucleotide data. *bioRxiv*.

Bouckaert, R.R., 2010. DensiTree: making sense of sets of phylogenetic trees. *Bioinformatics*, 26(10), pp.1372–1373. Available at: http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btq110 [Accessed August 29, 2016].

Broll, S. & Alexandersen, S., 1996. Investigation of the pathogenesis of transplacental transmission of Aleutian mink disease parvovirus in experimentally infected mink. *Journal of virology*, 70(3), pp.1455–66. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=189966&tool=pmcentrez&rendertype=abstract [Accessed May 21, 2015].

Brown, R.P. & Yang, Z., 2011. Rate variation and estimation of divergence times using strict and relaxed clocks. *BMC evolutionary biology*, 11(1), p.271. Available at: http://www.biomedcentral.com/1471-2148/11/271 [Accessed September 1, 2015].

Buratowski, S., 1994. The basics of basal transcription by RNA Polymerase II. *Cell*, 77, pp.1–3.

Burnham, K.P. & Anderson, D.R., 2002. *Model selection and multimodel inference: A practical-theoretic approach*, Springer-Verlag.

Bustin, S.A. et al., 2009. The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clinical Chemistry*, 55(4), pp.611–22. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19246619 [Accessed April 29, 2014].

Canuti, M. et al., 2016. Driving forces behind the evolution of the Aleutian mink disease parvovirus in the context of intensive farming. *Virus Evolution*, 2(1), p.vew004. Available at: http://ve.oxfordjournals.org/lookup/doi/10.1093/ve/vew004.

Cheng, F. et al., 2010. The capsid proteins of Aleutian mink disease virus activate caspases and are specifically cleaved during infection. *Journal of virology*, 84(6), pp.2687–96. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2826067&tool=pmcentrez&rendertype=abstract [Accessed March 20, 2015].

Cho, H.J. & Ingram, D.G., 1972. Antigen and antibody in Aleutian disease in mink. I. Precipitation reaction by agar-gel electrophoresis. *Journal of immunology (Baltimore, Md. : 1950)*, 108(2), pp.555–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/5049098 [Accessed October 3, 2016].

Chriél, M., 2000. Interpretation of test results in eradication programmes when multiple sampling is used. In *ISVEE 9: Proceedings of the 9th Symposium of the International Society for Veterinary Epidemiology and Economics*. Breckenrigde, CO, USA.

Christensen, J., Cotmore, S.F. & Tattersall, P., 1997. Parvovirus initiation factor PIF: a novel human DNA-binding factor which coordinately recognizes two ACGT motifs. *Journal*

of virology, 71(8), pp.5733–41. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=191825&tool=pmcentrez&rendertype=abstract [Accessed May 12, 2015].

Christensen, L.S. et al., 2011. Diversity and stability of Aleutian mink disease virus during bottleneck transitions resulting from eradication in domestic mink in Denmark. *Veterinary microbiology*, 149(1–2), pp.64–71. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21112164 [Accessed May 21, 2014].

Dam-Tuxen, R. et al., 2014. Diagnosing Aleutian mink disease infection by a new fully automated ELISA or by counter current immunoelectrophoresis: A comparison of sensitivity and specificity. *Journal of Virological Methods*, 199, pp.53–60.

Danecek, P. et al., 2011. The variant call format and VCFtools. *Bioinformatics (Oxford, England)*, 27(15), pp.2156–8. Available at: http://bioinformatics.oxfordjournals.org/content/27/15/2156 [Accessed July 10, 2014].

Darriba, D. et al., 2012. jModelTest 2: more models, new heuristics and parallel computing. *Nature methods*, 9(8), p.772. Available at: http://dx.doi.org/10.1038/nmeth.2109 [Accessed April 5, 2015].

Decaro, Nicola et al., 2012. 12. Parvovirus infections. In D. Gavier-Widén, J. P. Duff, & A. Meredith, eds. *Infectious Diseases of Wild Mammals and Birds in Europe*. Oxford, UK: Wiley-Blackwell, pp. 181–285.

Design, P. & Optimization, A., 2010. Real-Time PCR – The Basic Principles.

Dowgier, G. et al., 2016. A duplex real-time PCR assay based on TaqMan technology for simultaneous detection and differentiation of canine adenovirus types 1 and 2. *Journal of virological methods*. Available at: http://www.sciencedirect.com/science/article/pii/S0166093416300313 [Accessed April 5, 2016].

Drake, J.W., 1993. Rates of spontaneous mutation among RNA viruses. *Proceedings of the National Academy of Sciences of the United States of America*, 90(9), pp.4171–5. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8387212 [Accessed August 23, 2016].

Drummond, A.J. et al., 2007. A Rough Guide to BEAST 1 . 4. , pp.1–41.

Drummond, A.J. et al., 2005. Bayesian Coalescent Inference of Past Population Dynamics from Molecular Sequences. *Molecular Biology and Evolution*, 22(5), pp.1185–1192. Available at: http://mbe.oupjournals.org/cgi/doi/10.1093/molbev/msi103 [Accessed August 29, 2016].

Drummond, A.J. et al., 2012. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution*, 29(8), pp.1969–1973. Available at: http://mbe.oxfordjournals.org/cgi/doi/10.1093/molbev/mss075 [Accessed September 2, 2016].

Drummond, A.J. et al., 2002. Estimating Mutation Parameters, Population History and Genealogy Simultaneously From Temporally Spaced Sequence Data. *Genetics*, 161(3).

Drummond, A.J. et al., 2006. Relaxed phylogenetics and dating with confidence. *PLoS biology*, 4(5), p.e88. Available at: http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.0040088#pbio-0040088-b006 [Accessed July 9, 2014].

Drummond, A.J. & Bouckaert, R.R., 2015. Bayesian evolutionary analysis with BEAST 2. , p.249. Available at: https://books.google.com/books?id=olgZCgAAQBAJ&pgis=1 [Accessed September 1, 2015].

Duffy, S., Shackelton, L.A. & Holmes, E.C., 2008. Rates of evolutionary change in viruses: patterns and determinants. *Nature reviews. Genetics*, 9(4), pp.267–76. Available at: http://www.nature.com.globalproxy.cvt.dk/nrg/journal/v9/n4/abs/nrg2323.html [Accessed July 21, 2015].

Escobar-Gutiérrez, A. et al., 2012. Identification of hepatitis C virus transmission using a next-generation sequencing approach. *Journal of clinical microbiology*, 50(4), pp.1461–3. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3318530&tool=pmcentrez&rendertype=abstract [Accessed December 1, 2014].

Fahnø, U., 2014. Targeting the genetic complexity within adapting RNA virus populations. , (December).

Farid, A.H. & Ferns, L.., 2011. Aleutian mink disease virus infection may cause hair depigmentation. *Scientifur*, 35(September 2011), pp.55–59.

Farid, a H., 2013. Aleutian mink disease virus in furbearing mammals in Nova Scotia, Canada. *Acta veterinaria Scandinavica*, 55, p.10. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3602201&tool=pmcentrez&rendertype=abstract [Accessed May 13, 2014].

Fields, B.N., Knipe, D.M. & Howley, P.M., 2007. Fields Virology, 5th Edition. *Fields Virology*, 2, p.3177. Available at: http://www.loc.gov/catdir/toc/ecip072/2006032230.html.

Gilbert, J.A. et al., Chapter 12. , 733, pp.173–183.

Gilchrist, C.A. et al., 2015. Whole-Genome Sequencing in Outbreak Analysis. *Clinical Microbiology Reviews*, 28(3), pp.541–563. Available at: http://cmr.asm.org/lookup/doi/10.1128/CMR.00075-13 [Accessed August 16, 2016].

Gire, S.K. et al., 2014. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science*, 345(6202), pp.1369–72. Available at: http://www.sciencemag.org/content/345/6202/1369.abstract [Accessed August 29, 2014].

Gorm Pedersen, A., 2013. Course in Molecular Evolution. Available at: http://www.cbs.dtu.dk/courses/27615.mol/ [Accessed September 19, 2016].

Gottschalck, E. et al., 1994. Sequence comparison of the non-structural genes of four different types of Aleutian mink disease parvovirus indicates an unusual degree of variability. *Archives of Virology*, 138(3–4), pp.213–31. Available at: http://www.ncbi.nlm.nih.gov/pubmed/7998830 [Accessed May 27, 2014].

Grenfell, B.T. et al., 2004. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science (New York, N.Y.)*, 303(5656), pp.327–32. Available at: http://www.ncbi.nlm.nih.gov/pubmed/14726583 [Accessed August 29, 2016].

Hagberg, E.E., Larsen, L.E., Krarup, A., Pedersen, A.G., 2017. *Genetic analysis of the entire genome of Aleutian Mink Disease Virus determines its evolutionary rate and confirms bottleneck due to control program*,

Hagberg, E.E. et al., 2016. A fast and robust method for whole genome sequencing of the Aleutian Mink Disease Virus (AMDV) genome. *Journal of Virological Methods*. Available at: http://www.sciencedirect.com/science/article/pii/S0166093415300343.

Hagberg, E.E. et al., 2017. Evolutionary analysis of whole genome sequences from Aleutian Mink Disease Viruses confirms inter-farm transmission. *Journal of General Virology*.

Hasegawa, M., Kishino, H. & Yano, T., 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of molecular evolution*, 22(2), pp.160–74. Available at: http://www.ncbi.nlm.nih.gov/pubmed/3934395 [Accessed June 9, 2015].

Heath, T.A., Hedtke, S.M. & Hillis, D.M., 2008. Taxon sampling and the accuracy of phylogenetic analyses. *Journal of Systematics and Evolution*, 46(3), pp.239–257. Available at: http://www.plantsystematics.com [Accessed December 6,

2016].
Henn, M.R. et al., 2012. Whole genome deep sequencing of HIV-1 reveals the impact of early minor variants upon immune recognition during acute infection. C. M. Walker, ed. *PLoS pathogens*, 8(3), p.e1002529. Available at: http://dx.plos.org/10.1371/journal.ppat.1002529 [Accessed August 29, 2014].
Higuchi, R. et al., 1993. Kinetic PCR analysis: real-time monitoring of DNA amplification reactions. *Bio/technology (Nature Publishing Company)*, 11(9), pp.1026–30. Available at: http://www.ncbi.nlm.nih.gov/pubmed/7764001 [Accessed October 3, 2016].
Holder, M. & Lewis, P.O., 2003. Phylogeny estimation: traditional and Bayesian appraoches. , 4(April).
Huang, Q. et al., 2012. Internal polyadenylation of parvoviral precursor mRNA limits progeny virus production. *Virology*, 426(2), pp.167–77. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3294060&tool=pmcentrez&rendertype=abstract [Accessed May 21, 2014].
Huelsenbeck, J. & Rannala, B., 2004. Frequentist Properties of Bayesian Posterior Probabilities of Phylogenetic Trees Under Simple and Complex Substitution Models. *Systematic Biology*, 53(6), pp.904–913. Available at: http://sysbio.oxfordjournals.org/cgi/doi/10.1080/10635150490522629 [Accessed August 22, 2016].
Jakhesara, S.J. et al., 2014. Isolation and characterization of H9N2 influenza virus isolates from poultry respiratory disease outbreak. *SpringerPlus*, 3, p.196. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4004788&tool=pmcentrez&rendertype=abstract [Accessed May 26, 2014].
Jensen, T.H. et al., 2011. Implementation and validation of a sensitive PCR detection method in the eradication campaign against Aleutian mink disease virus. *Journal of virological methods*, 171(1), pp.81–5. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20951744 [Accessed May 21, 2014].
Jensen, T.H., Chriél, M. & Hansen, M.S., 2015. Progression of experimental chronic Aleutian mink disease virus infection. *Acta Veterinaria Scandinavica*, 58(1), p.35. Available at: http://actavetscand.biomedcentral.com/articles/10.1186/s13028-016-0214-7.
Katoh, K. & Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, 30(4), pp.772–80. Available at: http://mbe.oxfordjournals.org/content/30/4/772 [Accessed July 13, 2014].
Kearse, M. et al., 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics (Oxford, England)*, 28(12), pp.1647–9. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3371832&tool=pmcentrez&rendertype=abstract [Accessed July 10, 2014].
Knuuttila, A. et al., 2015. Aleutian mink disease virus in free-ranging mustelids in Finland - a cross-sectional epidemiological and phylogenetic study. *The Journal of general virology*, 96(Pt 6), pp.1423–35. Available at: http://jgv.sgmjournals.org.globalproxy.cvt.dk/content/journal/jgv/10.1099/vir.0.000081 [Accessed July 22, 2015].
Knuuttila, A. et al., 2009. Molecular epidemiology of Aleutian mink disease virus in Finland. *Veterinary Microbiology*, 133(3), pp.229–238.
Koboldt, D.C. et al., 2012. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research*, 22(3), pp.568–76. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22300766 [Accessed November 15, 2016].
Kryazhimskiy, S. & Plotkin, J.B., 2008. The population genetics of dN/dS. *PLoS genetics*, 4(12), p.e1000304. Available at: http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000304 [Accessed December 14, 2015].
Kühnert, D., Wu, C.-H. & Drummond, A.J., 2011. Phylogenetic and epidemic modeling of rapidly evolving infectious diseases. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases*, 11(8), pp.1825–41. Available at: http://www.sciencedirect.com/science/article/pii/S156713481100284X [Accessed December 17, 2015].
Kvisgaard, L.K. et al., 2013. A fast and robust method for full genome sequencing of Porcine Reproductive and Respiratory Syndrome Virus (PRRSV) Type 1 and Type 2. *J Virol Methods*, 193(2), pp.697–705. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23891870%5Cnhttp://ac.els-cdn.com/S0166093413002735/1-s2.0-S0166093413002735-main.pdf?_tid=2b32140c-d9f5-11e3-bf78-00000aacb360&acdnat=1399913456_ccd3c5666ff039c4dd4fc4a89b2d5abc.
Lander, E.S. et al., 2000. An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature*, 407(6803), pp.513–516. Available at: http://www.nature.com/doifinder/10.1038/35035083 [Accessed September 5, 2016].
Lee, W.-P. et al., 2014. MOSAIK: A Hash-Based Algorithm for Accurate Next-Generation Sequencing Short-Read Mapping C. K. Hsiao, ed. *PLoS ONE*, 9(3), p.e90581. Available at: http://dx.plos.org/10.1371/journal.pone.0090581 [Accessed September 6, 2016].
Leekitcharoenphon, P. et al., 2014. Evaluation of Whole Genome Sequencing for Outbreak Detection of Salmonella enterica J. A. Chabalgoity, ed. *PLoS ONE*, 9(2), p.e87991. Available at: http://dx.plos.org/10.1371/journal.pone.0087991 [Accessed August 15, 2016].
Leimann, A. et al., 2015. Molecular epidemiology of Aleutian mink disease virus (AMDV) in Estonia, and a global phylogeny of AMDV. *Virus research*, 199(2), pp.55–61. Available at: http://www.sciencedirect.com/science/article/pii/S0168170215000179 [Accessed January 26, 2015].
Li, H., 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Available at: http://arxiv.org/abs/1303.3997 [Accessed December 1, 2014].
Li, H. et al., 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16), pp.2078–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19505943 [Accessed February 4, 2017].
Li, L. et al., 2011. Novel amdovirus in gray foxes. *Emerging infectious diseases*, 17(10), pp.1876–8. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3310670&tool=pmcentrez&rendertype=abstract [Accessed November 4, 2014].
Li, L.M. et al., 2014. Genomic analysis of emerging pathogens: methods, application and future trends. *Genome Biology*, 15(11), p.541. Available at: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0541-9 [Accessed November 3, 2016].
Li, Y. et al., 2012. Genetic characterization of Aleutian mink disease viruses isolated in China. *Virus genes*, 45(1), pp.24–30. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22415541 [Accessed May 21, 2014].
Maclachlan, N.J., Dubovi, E.J. & Fenner, F., 2011. *Fenner's*

Veterinary Virology, Elsevier. Available at: http://www.sciencedirect.com/science/article/pii/B9780123751584000018 [Accessed May 7, 2015].
Martin, D.P. et al., 2015. RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evolution*, 1(1), pp.1–5. Available at: http://ve.oxfordjournals.org/cgi/doi/10.1093/ve/vev003.
Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1), pp.10–12. Available at: http://journal.embnet.org/index.php/embnetjournal/article/view/200.
McKenna, R. et al., 1999. Three-Dimensional Structure of Aleutian Mink Disease Parvovirus: Implications for Disease Pathogenicity. *J. Virol.*, 73(8), pp.6882–6891. Available at: http://jvi.asm.org/content/73/8/6882.abstract [Accessed March 19, 2015].
Melorose, J., Perroy, R. & Careas, S., 2009. *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing* 2nd ed. P. Lemey, M. Salemi, & A.-M. Vandamme, eds., Cambridge.
Metzker, M.L. et al., 2002. Molecular evidence of HIV-1 transmission in a criminal case. *Proceedings of the National Academy of Sciences of the United States of America*, 99(22), pp.14292–7. Available at: http://www.pnas.org/content/99/22/14292.full [Accessed August 3, 2015].
Metzker, M.L., 2010. Sequencing technologies — the next generation. *Nature Reviews Genetics*, 11(1), pp.31–46. Available at: http://www.nature.com/doifinder/10.1038/nrg2626 [Accessed January 14, 2017].
Morelli, M.J. et al., 2013. Evolution of foot-and-mouth disease virus intra-sample sequence diversity during serial transmission in bovine hosts. *Veterinary research*, 44(1), p.12. Available at: http://veterinaryresearch.biomedcentral.com/articles/10.1186/1297-9716-44-12 [Accessed May 17, 2016].
Mullis, K. et al., 1986. Specific Enzymatic Amplification of DNA In Vitro: The Polymerase Chain Reaction. *Cold Spring Harbor Symposia on Quantitative Biology*, 51(0), pp.263–273. Available at: http://symposium.cshlp.org/cgi/doi/10.1101/SQB.1986.051.01.032 [Accessed October 3, 2016].
Nelson, C.W. & Hughes, A.L., 2015. Within-host nucleotide diversity of virus populations: Insights from next-generation sequencing. *Infection, Genetics and Evolution*, 30, pp.1–7. Available at: http://dx.doi.org/10.1016/j.meegid.2014.11.026.
Nituch, L.A. et al., 2015. Aleutian mink disease virus in striped skunks (Mephitis mephitis): evidence for cross-species spillover. *Journal of wildlife diseases*, 51(2), pp.389–400. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25647590 [Accessed January 5, 2016].
Nituch, L. a. et al., 2012. Molecular epidemiology of Aleutian disease virus in free-ranging domestic, hybrid, and wild mink. *Evolutionary Applications*, 5(4), pp.330–340. Available at: http://doi.wiley.com/10.1111/j.1752-4571.2011.00224.x [Accessed May 21, 2014].
Oie, K.L. et al., 1996. The relationship between capsid protein (VP2) sequence and pathogenicity of Aleutian mink disease parvovirus (ADV): a possible role for raccoons in the transmission of ADV infections. *Journal of virology*, 70(2), pp.852–61. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=189888&tool=pmcentrez&rendertype=abstract.
Pamilo, P. & Nei, M., 1988. Relationships between gene trees and species trees. *Molecular biology and evolution*, 5(5), pp.568–83. Available at: http://www.ncbi.nlm.nih.gov/pubmed/3193878 [Accessed January 13, 2017].
Pedersen, A.G., 2012. seqconverter.
Pedersen, A.G., 2012. seqlib.py.
Pedersen, A.G., 2012. treerooter.py and treecutter.py.
Persson, S. et al., 2015. Aleutian mink disease virus in free-ranging mink from Sweden. *PloS one*, 10(3), p.e0122194. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4379071&tool=pmcentrez&rendertype=abstract [Accessed March 28, 2016].
du Plessis, L. & Stadler, T., 2015. Getting to the root of epidemic spread with phylodynamic analysis of genomic data. *Trends in Microbiology*, 23(7), pp.383–386. Available at: http://www.sciencedirect.com/science/article/pii/S096842X15001018 [Accessed July 3, 2015].
Pond, S. L., Frost, S.D.W., 2005. Datamonkey: Rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics*, 21(10), pp.2531–2533.
Posada, D., Crandall, K.A. & Holmes, E.C., 2002. Recombination in Evolutionary Genomics. *Annual Review of Genetics*, 36(1), pp.75–97. Available at: http://www.annualreviews.org/doi/10.1146/annurev.genet.36.040202.111115 [Accessed September 29, 2016].
Pybus, O.G., Rambaut, A. & Harvey, P.H., 2000. An Integrated Framework for the Inference of Viral Population History From Reconstructed Genealogies. *Genetics*, 155(3).
Qiu, J. et al., 2006. The transcription profile of Aleutian mink disease virus in CRFK cells is generated by alternative processing of pre-mRNAs produced from a single promoter. *Journal of virology*, 80(2), pp.654–62. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1346859&tool=pmcentrez&rendertype=abstract [Accessed November 4, 2014].
Quail, M. a et al., 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC genomics*, 13(1), p.341. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3431227&tool=pmcentrez&rendertype=abstract [Accessed April 28, 2014].
Quick, J. et al., 2015. Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of Salmonella. *Genome biology*, 16(1), p.114. Available at: http://genomebiology.com/2015/16/1/114 [Accessed June 1, 2015].
Quinlan, A.R. & Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, 26(6), pp.841–2. Available at: http://bioinformatics.oxfordjournals.org/content/26/6/841.short [Accessed July 9, 2014].
R core team, 2015. R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria.* Available at: https://www.r-project.org/.
Rambaut, A. et al., 2016. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evolution*, 2(1), pp.1–7.
Rodrigo A, G. & Felsenstein, J., 1999. Coalescent approaches to HIV population genetics. In *The Evolution of HIV*. pp. 233–272. Available at: https://books.google.dk/books?id=MGefyovUsgC&lpg=PA233&ots=naSsTCVdS7&dq=Coalescent approaches to hiv population genetics&hl=da&pg=PA234#v=onepage&q=Coalescent approaches to hiv population genetics&f=false.
Ronquist, F. et al., 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic biology*, 61(3), pp.539–42. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?art

id=3329765&tool=pmcentrez&rendertype=abstract [Accessed July 10, 2014].
Rosenberg, M.S. & Kumar, S., 2003. Taxon sampling, bioinformatics, and phylogenomics. *Systematic biology*, 52(1), pp.119–24. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2796430&tool=pmcentrez&rendertype=abstract [Accessed April 28, 2016].
Ryt-Hansen, P., Hagberg, E.E., et al., 2017. *Global Phylogenetic analysis of contemporary Aleutian Mink Disease Viruses (AMDVs)*,
Ryt-Hansen, P., Hjulsager, C.K., et al., 2017. *Outbreak investigation of Aleutian Mink Disease Virus (AMDV) using partial NS1 gene sequencing*,
Sainani, B.K., 2009. Evolution and HIV: Using Computational Phylogenetics to Close In On a Killer.
Sang, Y. et al., 2012. Phylogenetic analysis of the VP2 gene of Aleutian mink disease parvoviruses isolated from 2009 to 2011 in China. *Virus genes*, 45(1), pp.31–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22415542 [Accessed May 21, 2014].
Sanger, F., Nicklen, S. & Coulson, A.R., 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12), pp.5463–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/271968 [Accessed September 5, 2016].
Schmieder, R. & Edwards, R., 2011. Quality control and preprocessing of metagenomic datasets. *Bioinformatics {(Oxford,} England)*, 27(6), pp.863–864. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21278185.
Shackelton, L.A. et al., 2007. Comparative analysis reveals frequent recombination in the parvoviruses. *The Journal of general virology*, 88(Pt 12), pp.3294–301. Available at: http://jgv.microbiologyresearch.org/content/journal/jgv/10.1099/vir.0.83255-0#tab2 [Accessed February 1, 2016].
Snitkin, E.S. et al., 2012. Tracking a hospital outbreak of carbapenem-resistant Klebsiella pneumoniae with whole-genome sequencing. *Science translational medicine*, 4(148), p.148ra116. Available at: http://stm.sciencemag.org/content/4/148/148ra116.full [Accessed April 22, 2015].
Stadler, T. et al., 2013. Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proceedings of the National Academy of Sciences of the United States of America*, 110(1), pp.228–33. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3538216&tool=pmcentrez&rendertype=abstract [Accessed January 22, 2015].
Streck, A.F. et al., 2011. High rate of viral evolution in the capsid protein of porcine parvovirus. *Journal of General Virology*, 92(11), pp.2628–2636.
Strimmer, K. & Pybus, O.G., 2001. Exploring the demographic history of DNA sequences using the generalized skyline plot. *Molecular biology and evolution*, 18(12), pp.2298–305. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11719579 [Accessed June 28, 2016].
Themudo, G.E., Østergaard, J. & Ersbøll, A.K., 2011. Persistent spatial clusters of plasmacytosis among Danish mink farms. *Preventive veterinary medicine*, 102(1), pp.75–82. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21788091 [Accessed May 21, 2014].
Watson, S.J. et al., 2015. Molecular Epidemiology and Evolution of Influenza Viruses Circulating within European Swine between 2009 and 2013. *Journal of virology*, 89(19), pp.9920–31. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4577897&tool=pmcentrez&rendertype=abstract [Accessed December 1, 2015].
Willadsen, C.M., 2003. Rapport over forløbet af plasmacytose-epidemien blandt midt-/sydjyske- og fynske minkfarme i perioden 6. juni til 1. november 2002. *Dansk Pelsdyr Laboratorium, Glostrup*.
Wohl, S., Schaffner, S.F. & Sabeti, P.C., 2016. Genomic Analysis of Viral Outbreaks. *Annual Review of Virology*, 3(1), p.annurev-virology-110615-035747. Available at: http://www.annualreviews.org/doi/10.1146/annurev-virology-110615-035747 [Accessed August 16, 2016].
Xi, J. et al., 2016. Genetic characterization of the complete genome of an Aleutian mink disease virus isolated in north China. *Virus genes*. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27007772 [Accessed March 28, 2016].
Yang, Z. et al., 1996. Among-site rate variation and its impact on phylogenetic analyses. *Trends in ecology & evolution*, 11(9), pp.367–72. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21237881 [Accessed October 11, 2016].
Ye, J. et al., 2012. Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC bioinformatics*, 13, p.134. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3412702&tool=pmcentrez&rendertype=abstract [Accessed November 8, 2014].
Ypma, R.J.F. et al., 2013. Genetic data provide evidence for wind-mediated transmission of highly pathogenic avian influenza. *The Journal of infectious diseases*, 207(5), pp.730–5. Available at: http://jid.oxfordjournals.org/content/207/5/730.full [Accessed April 27, 2015].

APPENDIX

APPENDIX

I. Sample overview

II. Supplementary results

APPENDIX I

Sample overview

lbnr	country	region	city	cluster	fragm_size	samplingdate	avgcov
1	DK	NJ	Frederikshavn	Saeby	4369	2011-12-21	1561
2	DK	NJ	Ålbæk	Saeby	4369	2012-11-28	6187
8	DK	NJ	Frederikshavn	Saeby	4369	2014-11-17	3130
9	DK	NJ	Frederikshavn	Saeby	4369	2011-12-21	2524
10	DK	NJ	Frederikshavn	Saeby	4369	2011-12-21	1852
11	DK	NJ	Frederikshavn	Saeby	4369	2011-12-21	1617
12	DK	NJ	Ålbæk	Saeby	4369	2012-11-28	5121
13	DK	NJ	Ålbæk	Saeby	4369	2013-11-08	2348
16	DK	NJ	Brønderslev	Saeby	4369	2012-02-09	3850
17	DK	NJ	Hirtshals	Saeby	4369	2014-06-19	2118
18	DK	NJ	Hirtshals	Saeby	4369	2014-09-19	2203
19	DK	NJ	Hirtshals	Saeby	4369	2014-09-19	1974
20	DK	MJ	Sdr. Felding	Saeby	4369	2015-06-18	1056
21	DK	MJ	Sdr. Felding	Saeby	4369	2015-06-18	1227
22	DK	MJ	Sdr. Felding	Saeby	4369	2015-06-18	1305
23	DK	MJ	Sdr. Felding	Saeby	4369	2015-06-18	5514
24	DK	MJ	Sdr. Felding	Saeby	4369	2015-06-18	2821
25	DK	MJ	Sdr. Felding	Saeby	4369	2015-06-18	1166
27	DK	MJ	Holstebro	Saeby	4369	2014-11-18	3849
28	DK	MJ	Holstebro	Saeby	4369	2014-11-18	1865
29	DK	MJ	Holstebro	Saeby	4369	2014-11-18	2237
31	DK	NJ	Frederikshavn	Saeby	4369	2014-11-19	3267
32	DK	NJ	Frederikshavn	Saeby	4369	2014-11-19	1001
33	DK	NJ	Frederikshavn	Saeby	4369	2013-02-03	1551
34	DK	NJ	Frederikshavn	Saeby	4369	2013-02-03	735
35	DK	NJ	Frederikshavn	Saeby	4369	2014-02-10	1496
36	DK	NJ	Frederikshavn	Saeby	4369	2014-02-10	1339
37	DK	NJ	Frederikshavn	Saeby	4369	2014-02-10	765
40	DK	NJ	Bindslev	Saeby	4369	2014-11-11	1174
41	DK	NJ	Vodskov	Saeby	4369	2014-11-11	2337
43	DK	NJ	Vodskov	Saeby	4369	2014-02-24	3447
44	DK	NJ	Vodskov	Saeby	4369	2013-10-24	967
45	DK	NJ	Jerup	Saeby	4369	2015-02-27	927
46	DK	NJ	Jerup	Saeby	4369	2015-02-27	2375
47	DK	NJ	Jerup	Saeby	4369	2015-02-27	568
48	DK	NJ	Jerup	Saeby	4369	2011-12-15	782
50	DK	NJ	Jerup	Saeby	4369	2013-11-12	3523
51	DK	NJ	Jerup	Saeby	4369	2013-11-12	3236
52	DK	NJ	Jerup	Saeby	4369	2013-11-12	1272
53	DK	NJ	Jerup	Saeby	4369	2013-11-12	2890
54	DK	NJ	Brønderslev	Saeby	4369	2012-02-09	2171

55	DK	NJ	Brønderslev	Saeby	4369	2012-02-09	1552
56	DK	MJ	Holstebro	Saeby	4369	2014-06-27	2229
57	DK	MJ	Holstebro	Saeby	4369	2015-06-03	2218
58	DK	MJ	Holstebro	Saeby	4369	2015-06-03	2483
59	DK	MJ	Holstebro	Saeby	4369	2015-05-26	1496
60	DK	MJ	Holstebro	Saeby	4369	2015-05-26	1669
61	DK	MJ	Holstebro	Saeby	4369	2015-02-28	2589
62	DK	MJ	Struer	Saeby	4369	2014-12-12	2138
63	DK	MJ	Struer	Saeby	4369	2014-12-12	1681
64	DK	MJ	Struer	Saeby	4369	2014-12-12	1549
65	DK	MJ	Holstebro	Saeby	4369	2015-01-07	2157
66	DK	MJ	Holstebro	Saeby	4369	2015-01-07	1639
67	DK	MJ	Holstebro	Saeby	4369	2015-01-16	2330
68	DK	MJ	Holstebro	Saeby	4369	2015-01-16	1606
69	DK	MJ	Holstebro	Saeby	4369	2015-01-16	1421
70	DK	NJ	Hirtshals	Saeby	4369	2014-09-19	1587
71	DK	NJ	Hirtshals	Saeby	4369	2014-06-19	1341
72	DK	NJ	Hjallerup	Saeby	4369	2010-09-06	2540
73	DK	NJ	Hjallerup	Saeby	4369	2010-09-06	1427
75	NL	Gelderland	ukendt	Holstebro 2025/2016	4369	2010-04-17	3060
76	DK	NJ	Brønderslev	Saeby	4369	2009-02-13	3836
77	DK	NJ	Brønderslev	Saeby	4369	2009-02-13	1623
78	DK	NJ	Sindal	Saeby	4369	2009-02-13	1606
91	DK	NJ	Sindal	Saeby	4369	2009-02-13	1016
92	DK	NJ	Sindal	Saeby	4369	2009-02-13	4122
93	DK	NJ	Hjørring	Saeby	4369	2008-12-12	3441
94	DK	NJ	Hjørring	Saeby	4369	2008-12-12	777
95	DK	NJ	Tårs	Saeby	4369	2008-11-26	3577
96	DK	NJ	Tårs	Saeby	4369	2008-11-26	7655
97	DK	NJ	Tårs	Saeby	4369	2008-11-26	2335
98	DK	NJ	Tårs	Saeby	4369	2008-11-26	1425
99	DK	NJ	Tårs	Saeby	4369	2008-11-26	2648
100	DK	NJ	Tårs	Saeby	4369	2008-11-25	6661
101	DK	NJ	Hjallerup	Saeby	4369	2008-11-14	3356
102	DK	NJ	Hjallerup	Saeby	4369	2008-11-14	2774
103	DK	NJ	Jerslev	Saeby	4369	2008-11-14	1967
104	DK	NJ	Jerslev	Saeby	4369	2008-11-14	2275
105	DK	NJ	Vrå	Saeby	4369	2008-11-31	1588
107	DK	NJ	Sindal	Saeby	4369	2008-10-29	2884
108	DK	NJ	Sindal	Saeby	4369	2008-10-17	2466
109	DK	NJ	Sindal	Saeby	4369	2008-10-17	2286
110	DK	NJ	Hobro	Saeby	4369	2008-08-14	2081
111	DK	NJ	Hobro	Saeby	4369	2008-08-14	1701
112	DK	NJ	Hobro	Saeby	4369	2008-08-14	2075
113	DK	NJ	Hobro	Saeby	4369	2008-08-14	1382
114	DK	NJ	Hobro	Saeby	4369	2008-08-14	825
115	DK	NJ	Bindslev	Saeby	4369	2008-05-07	2576
116	DK	NJ	Bindslev	Saeby	4369	2008-05-07	2065
117	DK	NJ	Bindslev	Saeby	4369	2008-05-07	1596

119	DK	NJ	Sæby	Saeby	4369	2008-02-21	3032
120	DK	NJ	Sæby	Saeby	4369	2008-02-21	325
121	DK	NJ	Brønderslev	Saeby	4369	2008-02-18	99
122	DK	NJ	Sindal	Saeby	4369	2007-10-29	201
123	DK	NJ	Asaa	Saeby	4369	2007-10-10	301
124	DK	NJ	Asaa	Saeby	4369	2007-10-19	356
125	DK	NJ	Frederikshavn	Saeby	4369	2007-10-19	105
126	DK	NJ	Frederikshavn	Saeby	4369	2007-10-19	175
127	DK	NJ	Hjallerup	Saeby	4369	2007-09-13	454
128	DK	NJ	Hjallerup	Saeby	4369	2007-09-13	259
129	DK	NJ	Hjallerup	Saeby	4369	2007-09-13	127
130	DK	NJ	Jerup	Saeby	4369	2007-09-13	49
131	DK	NJ	Jerup	Saeby	4369	2007-09-13	80
132	DK	NJ	Jerup	Saeby	4369	2007-09-13	41
133	DK	NJ	Sæby	Saeby	4369	2007-05-25	71
134	DK	NJ	Bindslev	Saeby	4369	2007-03-07	69
135	DK	NJ	Frederikshavn	Saeby	4369	2007-02-09	107
136	DK	NJ	Frederikshavn	Saeby	4369	2007-02-09	100
137	DK	NJ	Sindal	Saeby	4369	2007-02-09	44
138	DK	NJ	Sindal	Saeby	4369	2007-02-09	75
139	DK	NJ	Sindal	Saeby	4369	2007-02-08	198
140	DK	NJ	Sindal	Saeby	4369	2007-02-08	45
141	DK	NJ	Sindal	Saeby	4369	2007-02-08	75
146	DK	MJ	Holstebro	Holstebro 2025/2016	4369	2015-09-17	7371
148	DK	MJ	Holstebro	Holstebro 2025/2016	4369	2015-10-06	2697
149	DK	MJ	Holstebro	Holstebro 2025/2016	4369	2015-10-06	1809
150	DK	MJ	Ulfborg	Holstebro 2025/2016	4369	2015-10-08	2961
151	DK	MJ	Skjern/Ringkøbing	Holstebro 2025/2016	4369	2015-10-12	2693
152	DK	MJ	Skjern/Ringkøbing	Holstebro 2025/2016	4369	2015-10-13	2860
155	DK	MJ	Skjern/Ringkøbing	Holstebro 2025/2016	4369	2015-10-13	2748
156	DK	MJ	Ulfborg	Holstebro 2025/2016	4369	2015-10-13	2017
157	DK	MJ	Ulfborg	Holstebro 2025/2016	4369	2015-10-13	2197
158	DK	MJ	Ulfborg	Holstebro 2025/2016	4369	2015-10-13	4829
160	DK	NJ	Hanstholm	Holstebro 2025/2016	4369	2015-10-16	1952
161	DK	NJ	Hanstholm	Holstebro 2025/2016	4369	2015-10-16	1927
162	DK	MJ	Løsning	Holstebro 2025/2016	4369	2015-10-21	3468
163	DK	MJ	Løsning	Holstebro 2025/2016	4369	2015-10-21	4375
164	DK	MJ	Holstebro	Holstebro 2025/2016	4369	2015-10-21	2834
165	DK	MJ	Holstebro	Holstebro 2025/2016	4369	2015-10-21	4096
166	DK	MJ	Holstebro	Holstebro 2025/2016	4369	2015-10-23	2444
167	DK	MJ	Holstebro	Holstebro 2025/2016	4369	2015-10-23	3013
168	DK	NJ	Års	Holstebro 2025/2016	4369	2015-10-29	6537
169	DK	NJ	Års	Holstebro 2025/2016	4369	2015-10-29	3771
171	DK	NJ	Frederikshavn	Saeby	4369	2013-12-04	929
172	DK	NJ	Hirtshals	Saeby	4369	2014-11-28	2525
173	DK	NJ	Hirtshals	Saeby	4369	2014-11-28	2014
175	DK	NJ	Vadum	Saeby	4369	2014-03-26	2801
176	DK	NJ	Vadum	Saeby	4369	2014-03-26	4338
177	DK	NJ	Vadum	Saeby	4369	2014-03-26	2030

179	DK	NJ	Hirtshals	Saeby	4369	2012-02-01	2154
180	DK	NJ	Hirtshals	Saeby	4369	2014-03-26	2967
181	DK	NJ	Hirtshals	Saeby	4369	2014-03-26	2546
182	DK	NJ	Hirtshals	Saeby	4369	2014-03-26	3257
185	DK	NJ	Hirtshals	Saeby	4369	2014-05-08	1283
187	DK	NJ	Hirtshals	Saeby	4369	2014-05-08	1438
189	DK	NJ	Tårs	Saeby	4369	2014-02-26	1514
192	DK	NJ	Frederikshavn	Saeby	4369	2013-11-28	1436
200	DK	NJ	Bindslev	Saeby	4369	2014-11-10	2263
204	DK	NJ	Åbrybro	Saeby	4369	2013-02-03	2592
205	DK	NJ	Åbrybro	Saeby	4369	2013-02-03	2458
206	DK	NJ	Åbrybro	Saeby	4369	2013-02-03	1594
211	DK	NJ	Frederikshavn	Saeby	4369	2014-11-25	2708
212	DK	NJ	Frederikshavn	Saeby	4369	2014-11-25	239
213	DK	NJ	Frederikshavn	Saeby	4369	2014-11-25	338
260	DK	SJ	Svinninge	Holstebro 2025/2016	4369	2015-12-04	729
288	DK	NJ	Brønderslev	Saeby	4369	2015-11-06	1638
290	DK	NJ	Storvorde	Holstebro 2025/2016	4369	2015-11-12	2407
291	DK	NJ	Storvorde	Holstebro 2025/2016	4369	2015-11-12	1559
292	DK	NJ	Storvorde	Holstebro 2025/2016	4369	2015-11-12	1534
293	DK	NJ	Storvorde	Holstebro 2025/2016	4369	2015-11-12	2221
295	DK	SD	Tommerup	Holstebro 2025/2016	4369	2015-12-28	1193
296	DK	SJ	Svinninge	Holstebro 2025/2016	4369	2015-12-28	1417
297	DK	SJ	Svinninge	Holstebro 2025/2016	4369	2015-12-28	1194
298	DK	SJ	Svinninge	Holstebro 2025/2016	4369	2015-12-28	1945
303	DK	SD	Tommerup	Holstebro 2025/2016	4369	2015-12-30	931
305	DK	SJ	Mørkøv	Holstebro 2025/2016	4369	2016-01-07	2157
306	DK	SJ	Mørkøv	Holstebro 2025/2016	4369	2016-01-07	1489
315	DK	NJ	Vodskov	Saeby	4369	2004-05-27	2801
316	DK	NJ	Læsø	Saeby	4369	2014-01-29	9320
319	DK	MJ	Nykøbing M.	Saeby	4369	2014-11-21	8347
320	DK	NJ	Dybvad	Saeby	4369	2014-02-24	23798
321	DK	NJ	Dybvad	Saeby	4369	2014-02-24	24082
322	DK	NJ	Dybvad	Saeby	4369	2014-02-24	11615
323	DK	NJ	Dybvad	Saeby	4369	2014-02-24	11903
324	DK	NJ	Dybvad	Saeby	4369	2014-02-24	277274
325	DK	NJ	Dybvad	Saeby	4369	2014-02-24	8565
327	DK	NJ	Vadum	Saeby	4369	2014-11-14	14211
328	DK	NJ	Vadum	Saeby	4369	2014-11-14	14238
329	DK	NJ	Vadum	Saeby	4369	2014-11-14	8015
331	DK	NJ	Vadum	Saeby	4369	2014-11-14	12768
332	DK	NJ	Vadum	Saeby	4369	2014-11-14	6008
333	DK	NJ	Vadum	Saeby	4369	2014-11-14	14170
342	DK	NJ	Vodskov	Saeby	4369	2014-11-21	1001
343	DK	NJ	Vodskov	Saeby	4369	2014-11-21	1496
344	DK	NJ	Vadum	Saeby	4369	2014-11-21	10959
345	DK	NJ	Vadum	Saeby	4369	2014-11-21	12912
346	DK	NJ	Vadum	Saeby	4369	2014-11-21	12627
347	DK	NJ	Hjallerup	Saeby	4369	2012-02-09	8734

406	PL	Wielkopolskie	Czerniejewo	Holstebro 2025/2016	4369	2016-02-07	1740
424	PL	Wielkopolskie	Czerniejewo	Holstebro 2025/2016	4369	2016-02-05	1935
425	PL	Wielkopolskie	Czerniejewo	Holstebro 2025/2016	4369	2016-02-05	1853
427	PL	Wielkopolskie	Czerniejewo	Holstebro 2025/2016	4369	2016-02-05	2412
452	PL	Żukowo	Kartuzy	Holstebro 2025/2016	4369	2016-02-04	1619
453	PL	Żukowo	Kartuzy	Holstebro 2025/2016	4369	2016-02-04	1078
454	PL	Żukowo	Kartuzy	Holstebro 2025/2016	4369	2016-02-04	1682
457	PL	Wielkopolskie	Broniszewice	Holstebro 2025/2016	4369	2016-02-10	1877
458	PL	Wielkopolskie	Broniszewice	Holstebro 2025/2016	4369	2016-02-10	2244
459	PL	Wielkopolskie	Broniszewice	Holstebro 2025/2016	4369	2016-02-10	1194
462	PL	Wielkopolskie	Lipowiec	Holstebro 2025/2016	4369	2016-02-10	2249
464	PL	Wielkopolskie	Lipowiec	Holstebro 2025/2016	4369	2016-02-10	1731
467	PL	Wielkopolskie	Starkowiec Piątkowski	Holstebro 2025/2016	4369	2016-02-09	1705
469	PL	Wielkopolskie	Starkowiec Piątkowski	Holstebro 2025/2016	4369	2016-02-09	1585
471	NL	Gelderland	Herveld	Holstebro 2025/2016	4369	2016-02-24	1992
472	NL	Gelderland	Herveld	Holstebro 2025/2016	4369	2016-02-24	1860
473	NL	Gelderland	Herveld	Holstebro 2025/2016	4369	2016-02-24	1607
481	NL	Nord-Brabant	Boeckel	Holstebro 2025/2016	4369	2016-02-24	1144
483	NL	Nord-Brabant	Boeckel	Holstebro 2025/2016	4369	2016-02-24	1283
3	DK	MJ	Højslev	Saeby	3198	2015-06-30	2094
4	DK	MJ	Højslev	Saeby	3198	2015-06-30	3708
5	DK	MJ	Holstebro	Saeby	3198	2014-06-27	1359
6	DK	MJ	Holstebro	Saeby	3198	2015-05-26	804
7	DK	MJ	Sdr. Felding	Saeby	3198	2015-06-18	4418
38	DK	NJ	Bindslev	Saeby	3198	2014-11-11	341
39	DK	NJ	Bindslev	Saeby	3198	2014-11-11	1040
42	DK	NJ	Vodskov	Saeby	3198	2014-11-11	1155
145	DK	MJ	Holstebro	Holstebro 2025/2016	3198	2015-09-17	3421
186	DK	NJ	Hirtshals	Saeby	3198	2014-05-08	1967
194	DK	NJ	Frederikshavn	Saeby	3198	2013-11-28	2288
195	DK	NJ	Ålbæk	Saeby	3198	2013-11-11	1396
196	DK	NJ	Ålbæk	Saeby	3198	2013-11-11	2592
197	DK	NJ	Ålbæk	Saeby	3198	2013-11-11	1188
198	DK	NJ	Bindslev	Saeby	3198	2014-11-10	1290
199	DK	NJ	Bindslev	Saeby	3198	2014-11-10	1277
214	DK	NJ	Vodskov	Saeby	3198	2013-10-24	821

APPENDIX II

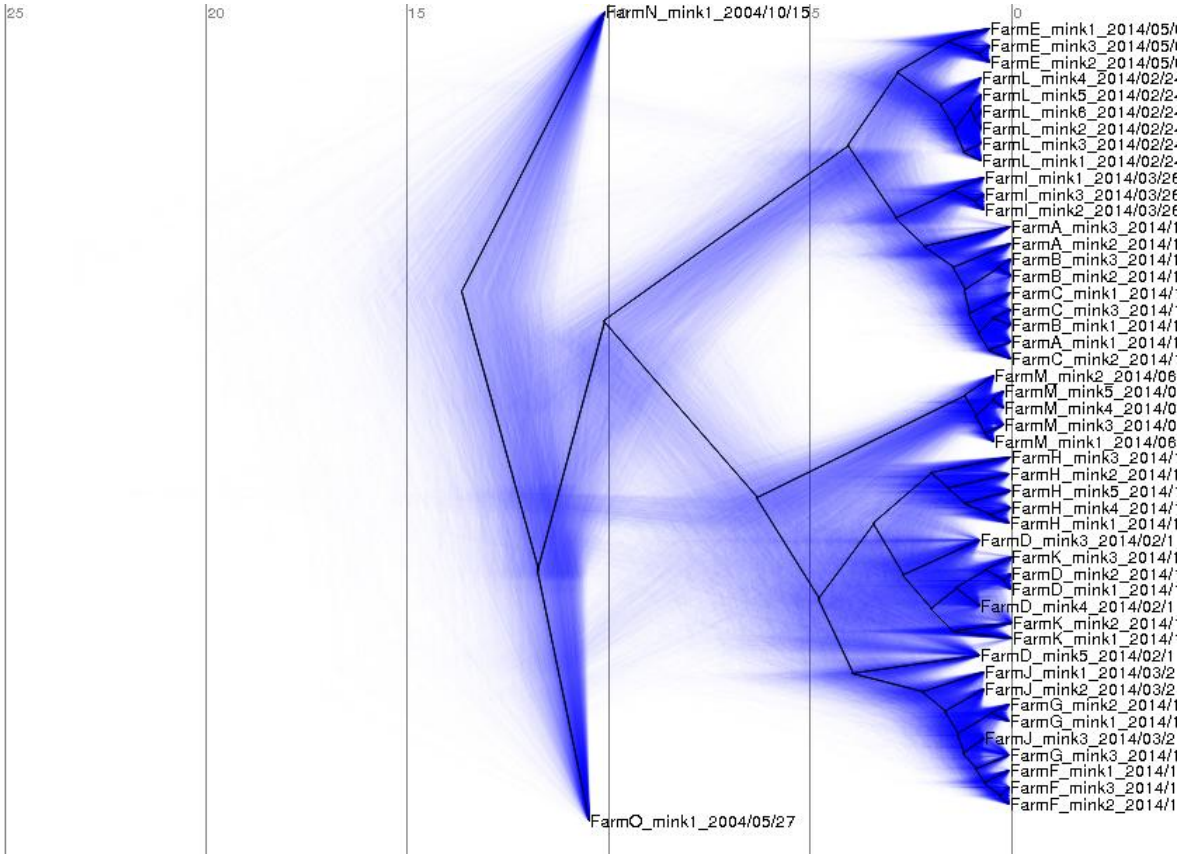


Figure 1. Alternative visualisation of the time-tree in Manucscript 2. The figure shows all the trees sampled with the MCMC during reconstruction of the time-tree in Manuscript 2 (figure 5), visualised on top of each other (blue) with the mcc-tree overlaid in black. The graph was created using DensiTree (Bouckaert, 2010).

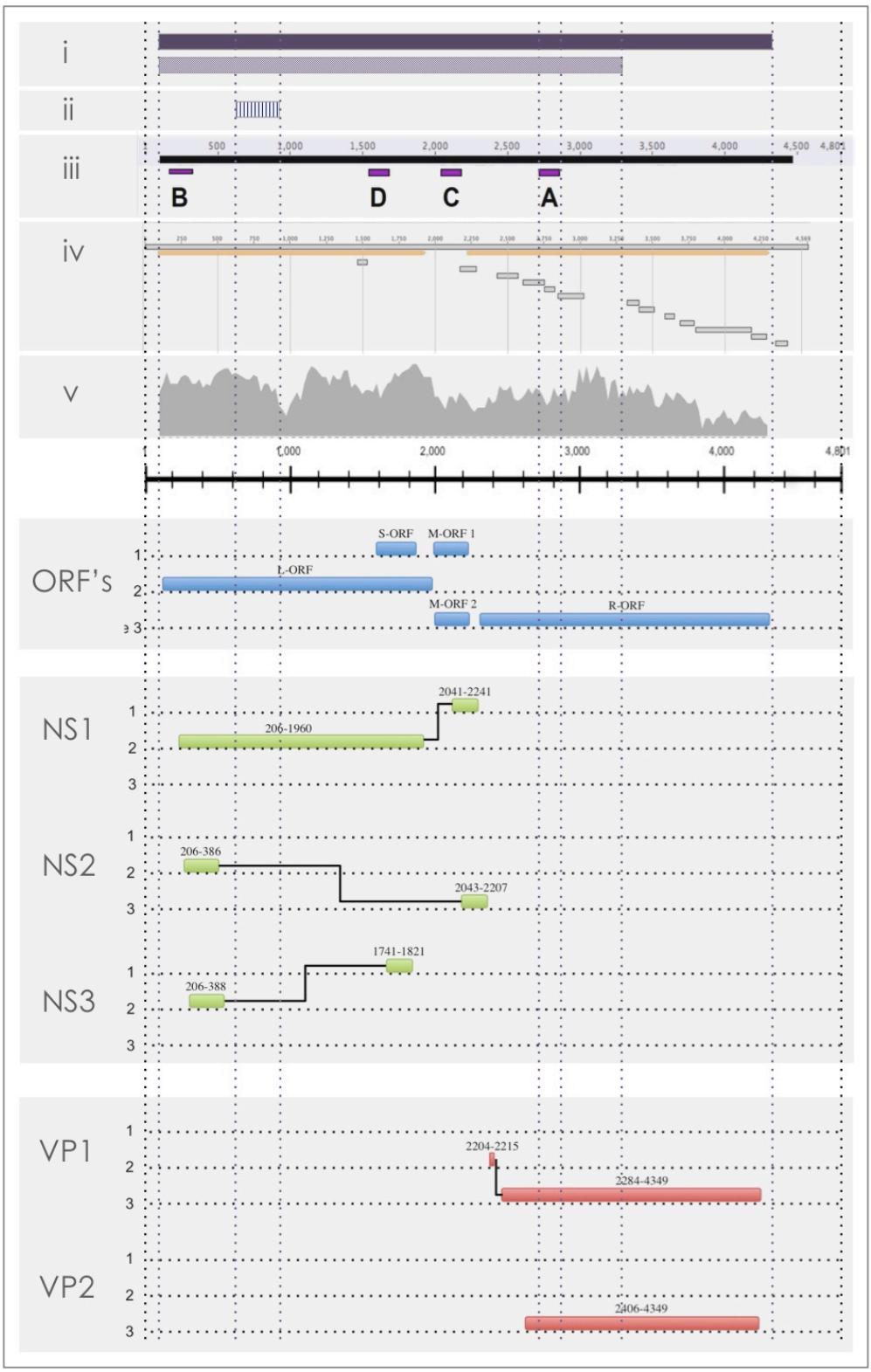


Figure 2. Extended genomic map of AMDV. The genomic map from Manuscript 1, indicating the localisation of the open reading frames (ORF), the major genes and their splicing. All numbering refers to AMDV-G (NC001662) nucleotide positions. Roman numbers: i) dark purple is the genomic region covered by the full-length PCR (nt. 98-4466), lighter purple is 'fragment A' (nt. 98-3295), ii) the region covered by the conventional PCR (nt. 578-951) by Jensen et al. (2011), iii) location of the real-time PCR primers from Manuscript 4, A is the region covered by final assay, iv) conserved genomic regions (Manuscript 3), and v) sliding-window analysis showing the phylogenetic resolution of 400-bp windows relative to WGS, the highest peaks are 40% (Manuscript 3).

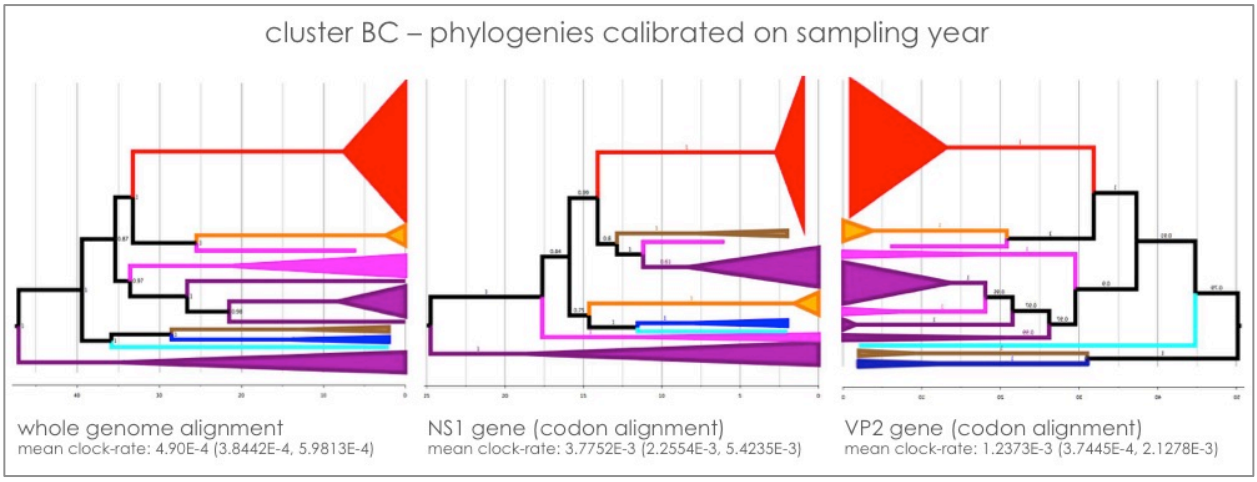


Figure 3. Topology comparison: WGS, NS1 and VP2. Holstebro sequences (red), Zealand (orange), Holland (pink), Poland (purple), Canada (blue), wildmink Canada (turquise), wildmink Bornholm (brown). Notice the red (Holstebro) sequence, which clusters separately based on the NS1 gene, while together with the other Holstebro sequences based on VP2. MCC trees from phylogenies constructed on each of the NS1 and VP2 genes for cluster BC using BEAST2 with an HKY-model, estimating the number of invariable sites and gamma-rate distribution from the data, calibration on sampling years, applying a strict molecular clock and an exponential (VP2) and constant (NS1) coalescent population model. The MCMC chains were run for 50M iterations.